STATISTICAL SCIENCE

Volume 41, Number 3 August 2026

The Infinitesimal Jackknife and Combinations of Models
. . . . *Indrayudh Ghosal, Yunzhe Zhou and Giles Hooker* 461

A Unified Theory of Exact Inference and Learning in Exponential Family Latent Variable Models . . . . *Sacha Sokoloski* 481

Jeffreys' Prior Penalty for High-Dimensional Logistic Regression: A Conjecture About Aggregate Bias . . . . *Ioannis Kosmidis and Patrick Zietkiewicz* 498

Posterior Ramifications of Prior Dependence Structures
. . . . *Luke Hagar and Nathaniel T. Stevens* 507

Evaluating Machine Learning Models in Nonstandard Settings: An Overview and New Findings . . . . *Roman Hornung, Malte Nalenz, Lennart Schneider, Andreas Bender, Ludwig Bothmann, Florian Dumpert, Bernd Bischl, Thomas Augustin and Anne-Laure Boulesteix* 524

A Review of Off-Policy Evaluation in Reinforcement Learning
. . . . *Masatoshi Uehara, Chengchun Shi and Nathan Kallus* 561

Gibbs Optimal Design of Experiments
. . . . *Antony M. Overstall, Jacinta Holloway-Brown and James M. McGree* 582

On Weighted Orthogonal Learners for Heterogeneous Treatment Effects
. . . . *Paweł Morzywołek, Johan Decruyenaere and Stijn Vansteelandt* 598

Restricted Maximum Likelihood Estimation in Generalized Linear Mixed Models
. . . . *Luca Maestrini, Francis K.C. Hui and Alan H. Welsh* 615

High-Probability Minimax Lower Bounds
. . . . *Tianyi Ma, Kabir A. Verchand and Richard J. Samworth* 637

The Fundamental Limits of Structure-Agnostic Functional Estimation
. . . . *Sivaraman Balakrishnan, Edward Kennedy and Larry Wasserman* 659

Sources of Uncertainty in Supervised Machine Learning—a Statisticians' View
. . . . *Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter and Göran Kauermann* 671

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 41, Number 3, August 2026. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

POSTMASTER: Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science **Volume 41, Number 3 (461–696) August 2026**

Volume 41 Number 3 August 2026

The Infinitesimal Jackknife and Combinations of Models
Indrayudh Ghosal, Yunzhe Zhou and Giles Hooker

A Unified Theory of Exact Inference and Learning in Exponential Family Latent Variable Models
Sacha Sokoloski

Jeffreys' Prior Penalty for High-Dimensional Logistic Regression: A Conjecture About Aggregate Bias
Ioannis Kosmidis and Patrick Zietkiewicz

Posterior Ramifications of Prior Dependence Structures
Luke Hagar and Nathaniel T. Stevens

Evaluating Machine Learning Models in Nonstandard Settings: An Overview and New Findings
Roman Hornung, Malte Nalenz, Lennart Schneider, Andreas Bender, Ludwig Bothmann, Florian Dumpert, Bernd Bischl, Thomas Augustin and Anne-Laure Boulesteix

A Review of Off-Policy Evaluation in Reinforcement Learning
Masatoshi Uehara, Chengchun Shi and Nathan Kallus

Gibbs Optimal Design of Experiments
Antony M. Overstall, Jacinta Holloway-Brown and James M. McGree

On Weighted Orthogonal Learners for Heterogeneous Treatment Effects
Paweł Morzywołek, Johan Decruyenaere and Stijn Vansteelandt

Restricted Maximum Likelihood Estimation in Generalized Linear Mixed Models
Luca Maestrini, Francis K.C. Hui and Alan H. Welsh

High-Probability Minimax Lower Bounds
Tianyi Ma, Kabir A. Verchand and Richard J. Samworth

The Fundamental Limits of Structure-Agnostic Functional Estimation
Sivaraman Balakrishnan, Edward Kennedy and Larry Wasserman

Sources of Uncertainty in Supervised Machine Learning—a Statisticians' View
Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter and Göran Kauermann

Statistical Science
2026, Vol. 41, No. 3, 461–480
https://doi.org/10.1214/24-STS959

The Infinitesimal Jackknife and Combinations of Models

Indrayudh Ghosal, Yunzhe Zhou and Giles Hooker

Abstract. The Infinitesimal Jackknife is a general method for estimating variances of parametric models and, more recently, also for some ensemble methods. In this paper we extend the Infinitesimal Jackknife to estimate the covariance between any two models. This can be used to quantify uncertainty for combinations of models or to construct test statistics for comparing different models or ensembles of models fitted using the same training dataset. Specific examples in this paper use boosted combinations of models, like random forests and M-estimators. We also investigate its application on neural networks and ensembles of XGBoost models. We illustrate the efficacy of variance estimates through extensive simulations and its application to the Beijing Housing data and demonstrate the theoretical consistency of the Infinitesimal Jackknife covariance estimate. The code is publicly available at https://github.com/yunzhe-zhou/IJ_ComModels.

Key words and phrases: Infinitesimal Jackknife, M-estimator, ensemble, kernel, V-statistics, random forests.

REFERENCES

[1] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. et al. (2016). Tensorflow: A system for large-scale machine learning. In 12*th* {*USENIX*} *Symposium on Operating Systems Design and Implementation* ({*OSDI*} 16) 265–283.

[2] ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

[3] ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

[4] BOTTOU, L. et al. (1991). Stochastic gradient learning in neural networks. *Proc. Neuro-Nımes* **91** 12.

[5] BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

[6] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

[7] CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.

[8] EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA. MR0659849

[9] EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671 https://doi.org/10.1080/01621459.2013.823775

[10] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

[11] GHOSAL, I. and HOOKER, G. (2021). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *J. Comput. Graph. Statist.* **30** 493–502. MR4270519 https://doi.org/10.1080/10618600.2020.1820345

[12] GHOSAL, I. and HOOKER, G. (2021). Generalised Boosted Forests. ArXiv preprint. Available at arXiv:2102.12561.

[13] GHOSAL, I., ZHOU, Y. and HOOKER, G. (2026). Supplement to "The Infinitesimal Jackknife and Combinations of Models." https://doi.org/10.1214/24-STS959SUPP

[14] HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. *Springer Series in Statistics*. Springer, New York. MR1461272 https://doi.org/10.1007/978-1-4757-2722-7

[15] HAYASHI, F. (2000). Extremum estimators. *Econometrics*. Princeton University Press, Princeton, New Jersey.

[16] HUANG, Z., LAM, H. and ZHANG, H. (2021). Quantifying Epistemic Uncertainty in Deep Learning. ArXiv preprint. Available at arXiv:2110.12122.

[17] JAECKEL, L. A. (1972). *The Infinitesimal Jackknife*. Bell Telephone Laboratories.

[18] KINGMA, D. P. and BA, J. (2014). Adam: a method for stochastic optimization. ArXiv preprint. Available at arXiv:1412.6980.

[19] LU, B. and HARDIN, J. (2021). A unified framework for random forest prediction error estimation. *J. Mach. Learn. Res.* **22** Paper No. 8. MR4253701

Indrayudh Ghosal is Ph.D., Department of Statistics and Data Science, Cornell University, Ithaca, New York 14853, USA (e-mail: indrayudhghosal1992@gmail.com). Yunzhe Zhou is Ph.D., Department of Biostatistics, University of California, Berkeley, Berkeley, California 94720-7360, USA (e-mail: ztzyz615@berkeley.edu). Giles Hooker is Professor, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: ghooker@wharton.upenn.edu).

[20] MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

[21] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26. MR3491120

[22] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

[23] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

[24] PENG, W., COLEMAN, T. and MENTCH, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electron. J. Stat.* **16** 232–292. MR4359361 https://doi.org/10.1214/21-ejs1958

[25] RUIQURM (2018). Kaggle. House price in Beijing.

[26] SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (2006). *Variance Components*. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. Reprint of the 1992 original, Wiley-Interscience Paperback Series. MR2298115

[27] SUN, R.-Y. (2020). Optimization for deep learning: An overview. *J. Oper. Res. Soc. China* **8** 249–294. MR4114365 https://doi.org/10.1007/s40305-020-00309-6

[28] SUN, Y., APLEY, D. W. and STAUM, J. (2011). Efficient nested simulation for estimating the variance of a conditional expectation. *Oper. Res.* **59** 998–1007. MR2844419 https://doi.org/10.1287/opre.1110.0932

[29] TOMITA, T. M., BROWNE, J., SHEN, C., CHUNG, J., PATSOLIC, J. L., FALK, B., PRIEBE, C. E., YIM, J., BURNS, R. et al. (2020). Sparse projection oblique randomer forests. *J. Mach. Learn. Res.* **21** 1–39.

[30] TUKEY, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Stat.* **29** 614.

[31] VAN DER VAART, A. W. and WELLNER, J. A. (2023). *Weak Convergence and Empirical Processes—with Applications to Statistics*. *Springer Series in Statistics*. Springer, Cham. MR4628026 https://doi.org/10.1007/978-3-031-29040-4

[32] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[33] WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15** 1625–1651. MR3225243

[34] WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372. MR0185765

[35] WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** 1–17.

[36] XU, T., ZHU, R. and SHAO, X. (2022). On Variance Estimation of Random Forests. ArXiv preprint. Available at arXiv:2202.09008.

[37] ZHOU, Y. and HOOKER, G. (2022). Boulevard: Regularized stochastic gradient boosted trees and their limiting distribution. *J. Mach. Learn. Res.* **23** Paper No. 183. MR4577136

[38] ZHOU, Z., MENTCH, L. and HOOKER, G. (2021). V-statistics and variance estimation. *J. Mach. Learn. Res.* **22** Paper No. 287. MR4353066

Statistical Science
2026, Vol. 41, No. 3, 481–497
https://doi.org/10.1214/25-STS979

A Unified Theory of Exact Inference and Learning in Exponential Family Latent Variable Models

Sacha Sokoloski

Abstract. Bayes' rule describes how to infer posterior beliefs about latent variables given observations, and inference is a critical step in learning algorithms for latent variable models (LVMs). Although there are exact algorithms for inference and learning for certain LVMs, such as linear Gaussian models and mixture models, researchers must typically develop approximate inference and learning algorithms when applying novel LVMs. Here, we study the line that separates LVMs that rely on approximation schemes from those that do not, and develop a general theory of exponential family LVMs for which inference and learning may be implemented exactly. First, under mild assumptions about the exponential family form of the LVM, we derive a necessary and sufficient constraint on the parameters of the LVM under which the prior and posterior over the latent variables are in the same exponential family. We then show that a variety of well-known and novel models indeed have this constrained, exponential family form. Finally, we derive generalized inference and learning algorithms for these LVMs, and demonstrate them with a variety of examples. Our unified perspective facilitates both understanding and implementing exact inference and learning algorithms for a wide variety of models, and may guide researchers in the discovery of new models that avoid unnecessary approximations.

Key words and phrases: Bayesian inference, maximum likelihood, exponential families, latent variable models.

REFERENCES

[1] ACKLEY, D. H., HINTON, G. E. and SEJNOWSKI, T. J. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9** 147–169.

[2] AMARI, S. and NAGAOKA, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Amer. Math. Soc., Providence, RI. Translated from the 1993 Japanese original by Daishi Harada. MR1800071 https://doi.org/10.1090/mmono/191

[3] ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1993). Conjugate exponential family priors for exponential family likelihoods. *Statistics* **25** 71–77. Publisher: Taylor & Francis. MR1268394 https://doi.org/10.1080/02331889308802432

[4] ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001). Conditionally specified distributions: An introduction (with comments and a rejoinder by the authors). *Statist. Sci.* **16** 249–274. Publisher: Institute of Mathematical Statistics. MR1874154 https://doi.org/10.1214/ss/1009213728

[5] ARNOLD, B. C. and PRESS, S. J. (1989). Compatible conditional distributions. *J. Amer. Statist. Assoc.* **84** 152–156. MR0999673

[6] BECK, J., LATHAM, P. and POUGET, A. (2011). Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31** 15310–15319.

[7] BECK, J., POUGET, A. and HELLER, K. A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) **25** 3059–3067. Curran Associates, Red Hook.

[8] BENGIO, Y. and DELALLEAU, O. (2009). Justifying and generalizing contrastive divergence. *Neural Comput.* **21** 1601–1621. MR2527797 https://doi.org/10.1162/neco.2008.11-07-647

[9] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B, Methodol.* **36** 192–236. MR0373208

[10] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

[11] BOULANGER-LEWANDOWSKI, N., BENGIO, Y. and VINCENT, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the* 29*th International*

Sacha Sokoloski is Group Leader, Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany (e-mail: sacha.sokoloski@mailbox.org).

Coference on International Conference on Machine Learning. *ICML*'12 1881–1888. Omnipress, USA.

[12] CUNNINGHAM, J. P. and YU, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17** 1500–1509. https://doi.org/10.1038/nn.3776

[13] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B, Methodol.* **39** 1–22.

[14] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. Publisher: Institute of Mathematical Statistics. MR0520238

[15] DURSTEWITZ, D. (2017). A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLoS Comput. Biol.* **13** e1005542. Publisher: Public Library of Science. https://doi.org/10.1371/journal.pcbi.1005542

[16] GERWINN, S., BERENS, P. and BETHGE, M. (2009). *A Joint Maximum-Entropy Model for Binary Neural Population Patterns and Continuous Signals*. *Advances in Neural Information Processing Systems* **22**. Curran Associates, Red Hook.

[17] HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800.

[18] HINTON, G. E., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18** 1527–1554. MR2224485 https://doi.org/10.1162/neco.2006.18.7.1527

[19] HO, J., JAIN, A. and ABBEEL, P. (2020). Denoising diffusion probabilistic models. In *In Advances in Neural Information Processing Systems* **33** 6840–6851. Curran Associates, Red Hook.

[20] JAYNES, E. T. (2003). *Probability Theory the Logic of Science*. Cambridge Univ. Press, Cambridge. Edited and with a foreword by G. Larry Bretthorst. MR1992316 https://doi.org/10.1017/CBO9780511790423

[21] KINGMA, D. and BA, J. (2014). Adam: A method for stochastic optimization. arXiv preprint. Available at arXiv:1412.6980.

[22] KINGMA, D. P. and WELLING, M. (2019). An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12** 307–392. https://doi.org/10.1561/2200000056

[23] KREFL, D., CARRAZZA, S., HAGHIGHAT, B. and KAHLEN, J. (2020). Riemann-Theta Boltzmann machine. *Neurocomputing* **388** 334–345. https://doi.org/10.1016/j.neucom.2020.01.011

[24] LAKE, B. M., SALAKHUTDINOV, R. and TENENBAUM, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science* **350** 1332–1338. MR3445697 https://doi.org/10.1126/science.aab3050

[25] LANGDON, C., GENKIN, M. and ENGEL, T. A. (2023). A unifying perspective on neural manifolds and circuits for cognition. In *Nature Reviews Neuroscience* **24** 363–377. Nature Publishing Group. https://doi.org/10.1038/s41583-023-00693-x

[26] LANGE, R. D., SHIVKUMAR, S., CHATTORAJ, A. and HAEFNER, R. M. (2023). Bayesian encoding and decoding as distinct perspectives on neural coding. In *Nature Neuroscience* **26** 2063–2072. Nature Publishing Group. https://doi.org/10.1038/s41593-023-01458-6

[27] MA, W. J., BECK, J., LATHAM, P. and POUGET, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9** 1432–1438. https://doi.org/10.1038/nn1790

[28] MARTIN, G. M., FRAZIER, D. T. and ROBERT, C. P. (2024). Approximating Bayes in the 21st century. *Statist. Sci.* **39** 20–45. Publisher: Institute of Mathematical Statistics. MR4718525 https://doi.org/10.1214/22-sts875

[29] MURPHY, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press, Cambridge.

[30] ROWEIS, S. and GHAHRAMANI, Z. (1999). A unifying review of linear Gaussian models. *Neural Comput.* **11** 305–345. https://doi.org/10.1162/089976699300016674

[31] SALAKHUTDINOV, R. and HINTON, G. (2012). An efficient learning procedure for deep Boltzmann machines. *Neural Comput.* **24** 1967–2006. MR2978160 https://doi.org/10.1162/NECO_a_00311

[32] SALAKHUTDINOV, R., TENENBAUM, J. B. and TORRALBA, A. (2013). Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1958–1971. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2012.269

[33] SÄRKKÄ, S. (2013). *Bayesian Filtering and Smoothing*. *Institute of Mathematical Statistics Textbooks* **3**. Cambridge Univ. Press, Cambridge. MR3154309 https://doi.org/10.1017/CBO9781139344203

[34] SCHNEIDMAN, E. (2016). Towards the design principles of neural population codes. *Curr. Opin. Neurobiol.* **37** 133–140. https://doi.org/10.1016/j.conb.2016.03.001

[35] SHEKHOVTSOV, A., SCHLESINGER, D. and FLACH, B. (2021). VAE Approximation Error: ELBO and Exponential Families.

[36] SHIVKUMAR, S., LANGE, R., CHATTORAJ, A. and HAEFNER, R. (2018). *A Probabilistic Population Code Based on Neural Samples*. *Advances in Neural Information Processing Systems* **31**. Curran Associates, Red Hook.

[37] SHMUELI, G., MINKA, T. P., KADANE, J. B., BORLE, S. and BOATWRIGHT, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *J. R. Stat. Soc., Ser. C* **54** 127–142. MR2134602 https://doi.org/10.1111/j.1467-9876.2005.00474.x

[38] SHORE, J. E. and JOHNSON, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **26** 26–37. MR0560389 https://doi.org/10.1109/TIT.1980.1056144

[39] SMOLENSKY, P. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory Technical Report, Colorado Univ. at Boulder Dept. Computer Science. Section: Technical Reports.

[40] SOKOLOSKI, S. (2026). Supplement to "A unified theory of exact inference and learning in exponential family latent variable models." https://doi.org/10.1214/25-STS979SUPP

[41] SOKOLOSKI, S., ASCHNER, A. and COEN-CAGLI, R. (2021). Modelling the neural code in large populations of correlated neurons. *eLife* **10** e64615. Publisher: eLife Sciences Publications, Ltd. https://doi.org/10.7554/eLife.64615

[42] STEVENSON, I. H. (2016). Flexible models for spike count data with both over- and under-dispersion. *J. Comput. Neurosci.* **41** 29–43. MR3517604 https://doi.org/10.1007/s10827-016-0603-y

[43] SUTSKEVER, I. and TIELEMAN, T. (2010). On the convergence properties of contrastive divergence. In *International Conference on Artificial Intelligence and Statistics* 789–795.

[44] TANSEY, W., PADILLA, O. H. M., SUGGALA, A. S. and RAVIKUMAR, P. (2015). Vector-space Markov random fields via exponential families. In *PMLR* 684–692.

[45] TAYLOR, G. W., HINTON, G. E. and ROWEIS, S. T. (2011). Two distributed-state models for generating high-dimensional time series. *J. Mach. Learn. Res.* **12** 1025–1068. MR2786917

[46] VAHDAT, A., ANDRIYASH, E. and MACREADY, W. (2020). Undirected graphical models as approximate posteriors. In *Proceedings of the 37th International Conference on Machine Learning* 9680–9689. PMLR. ISSN: 2640-3498.

[47] VÉRTES, E. and SAHANI, M. (2018). Flexible and accurate inference and learning for deep generative models. In *Advances*

in Neural Information Processing Systems (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) **31** 4166–4175. Curran Associates, Red Hook.

[48] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found*. *Trends Mach*. *Learn*. **1** 1–305.

[49] WELLING, M., ROSEN-ZVI, M. and HINTON, G. E. (2005). Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems* (L. K. Saul, Y. Weiss and L. Bottou, eds.) **17** 1481–1488. MIT Press, Cambridge, MA.

[50] YANG, E., ALLEN, G., LIU, Z. and RAVIKUMAR, P. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* **25**. Curran Associates, Red Hook.

[51] YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach*. *Learn*. *Res*. **16** 3813–3847. MR3450553

Statistical Science
2026, Vol. 41, No. 3, 498–506
https://doi.org/10.1214/25-STS980

Jeffreys' Prior Penalty for High-Dimensional Logistic Regression: A Conjecture About Aggregate Bias

Ioannis Kosmidis and Patrick Zietkiewicz

Abstract. Firth (1993, Biometrika) shows that the maximum Jeffreys' prior penalized likelihood estimator in logistic regression has asymptotic bias decreasing with the square of the number of observations when the number of parameters is fixed, which is an order faster than the typical rate from maximum likelihood. The widespread use of that estimator in applied work is supported by the results in Kosmidis and Firth (2021, Biometrika), who show that it takes finite values, even in cases where the maximum likelihood estimate does not exist. Kosmidis and Firth (2021, Biometrika) also provide empirical evidence that the estimator has good bias properties in high-dimensional settings where the number of parameters grows asymptotically linearly but slower than the number of observations. We design and carry out a large-scale computer experiment covering a wide range of such high-dimensional settings and produce strong empirical evidence for a simple rescaling of the maximum Jeffreys' prior penalized likelihood estimator that delivers high accuracy in signal recovery, in terms of aggregate bias, in the presence of an intercept parameter. The rescaled estimator is effective even in cases where estimates from maximum likelihood and other recently proposed corrective methods based on approximate message passing do not exist.

Key words and phrases: Mean bias reduction, approximate message passing, infinite estimates.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

CANDÈS, E. J. and SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* **48** 27–42. MR4065151 https://doi.org/10.1214/18-AOS1789

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212 https://doi.org/10.1093/biomet/80.1.27

HEINZE, G., PLONER, M., JIRICKA, L. and STEINER, G. (2023). logistf: Firth's Bias-Reduced Logistic Regression R package version 1.25.0. https://doi.org/10.32614/CRAN.package.logistf

KONIS, K. (2007). Linear programming algorithms for detecting separated data in binary logistic regression models Ph.D. thesis, Oxford Univ. Press, London.

KOSMIDIS, I. (2023). brglm2: Bias Reduction in Generalized Linear Models R package version 0.9. https://doi.org/10.32614/CRAN.package.brglm2

KOSMIDIS, I. and FIRTH, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108** 71–82. MR4226190 https://doi.org/10.1093/biomet/asaa052

KOSMIDIS, I., KENNE PAGUI, E. C. and SARTORI, N. (2020). Mean and median bias reduction in generalized linear models. *Stat. Comput.* **30** 43–59. MR4057470 https://doi.org/10.1007/s11222-019-09860-6

KOSMIDIS, I., SCHUMACHER, D. and SCHWENDINGER, F. (2022). detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates R package version 0.3. https://doi.org/10.32614/CRAN.package.detectseparation

KOSMIDIS, I. and ZIETKIEWICZ, P. (2026). Supplement to "Jeffreys' prior penalty for high-dimensional logistic regression: A conjecture about aggregate bias." https://doi.org/10.1214/25-STS980SUPP

MAK, S. (2021). minimaxdesign: Minimax and Minimax Projection Designs R package version 0.1.5. Available at https://cran.r-project.org/package=minimaxdesign.

MAK, S. and JOSEPH, V. R. (2018). Minimax and minimax projection designs using clustering. *J. Comput. Graph. Statist.* **27** 166–178.

Ioannis Kosmidis is the Professor of Statistics, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK (e-mail: ioannis.kosmidis@warwick.ac.uk). Patrick Zietkiewicz is a PhD student, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK (e-mail: patrick.zietkiewicz@warwick.ac.uk).

MR3788310 https://doi.org/10.1080/10618600.2017.1302881

McCullagh, P. (1987). *Tensor Methods in Statistics*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0907286

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1201/9780203753736

Salehi, F., Abbasi, E. and Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems* **1075** 12005–12015. Curran Associates, Red Hook, NY.

Sur, P. (2019). *A Modern Maximum Likelihood Theory for High-Dimensional Logistic Regression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University. MR4197622

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447. MR0375592 https://doi.org/10.1093/biomet/61.3.439

Yadlowsky, S., Yun, T., McLean, C. Y. and D'Amour, A. (2021). SLOE: A faster method for statistical inference in high-dimensional logistic regression. *Adv*. *Neural Inf*. *Process*. *Syst*. **34** 29517–29528.

Zhao, Q., Sur, P. and Candès, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28** 1835–1861. MR4411513 https://doi.org/10.3150/21-bej1401

Statistical Science
2026, Vol. 41, No. 3, 507–523
https://doi.org/10.1214/25-STS981

Posterior Ramifications of Prior Dependence Structures

Luke Hagar and Nathaniel T. Stevens

Abstract. Prior elicitation methods for Bayesian analyses transfigure prior information into quantifiable prior distributions. Recently, methods that leverage copulas have been proposed to accommodate more flexible dependence structures when eliciting multivariate priors. We show that the posterior cannot retain many of these flexible prior dependence structures in large-sample settings, and we emphasize that it is our responsibility as statisticians to communicate this to practitioners. We therefore overview objectives for prior specification that guide conversations between statisticians and practitioners to promote alignment between the flexibility in the prior dependence structure and the objectives for posterior analysis. Because correctly specifying the dependence structure a priori can be difficult, we consider how the choice of prior copula impacts the posterior distribution in terms of asymptotic convergence of the posterior mode. Our resulting recommendations clarify when it is useful to elicit intricate prior dependence structures and when it is not.

Key words and phrases: Copulas, credible sets, prior elicitation, the Bernstein–von Mises theorem.

REFERENCES

[1] BARONE, R. and DALLA VALLE, L. (2023). Bayesian nonparametric modeling of conditional multidimensional dependence structures. *J. Comput. Graph. Statist.* **32** 1361–1370. MR4669252 https://doi.org/10.1080/10618600.2023.2173604

[2] BEDFORD, T. and COOKE, R. M. (2002). Vines—a new graphical model for dependent random variables. *Ann. Statist.* **30** 1031–1068. MR1926167 https://doi.org/10.1214/aos/1031689016

[3] BERRY, S. M., CARLIN, B. P., LEE, J. J. and MÜLLER, P. (2011). *Bayesian Adaptive Methods for Clinical Trials. Chapman & Hall/CRC Biostatistics Series* **38**. CRC Press, Boca Raton, FL. MR2723582

[4] CARLIN, B. P. and LOUIS, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Springer, Berlin.

[5] CASELLA, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118 https://doi.org/10.2307/2682801

[6] CASEMENT, C. J. and KAHLE, D. J. (2018). Graphical prior elicitation in univariate models. *Comm. Statist. Simulation Comput.* **47** 2906–2924. MR3874066 https://doi.org/10.1080/03610918.2017.1361981

[7] CASEMENT, C. J. and KAHLE, D. J. (2023). The phoropter method: A stochastic graphical procedure for prior elicitation in univariate data models. *J. Korean Statist. Soc.* **52** 60–82. MR4552604 https://doi.org/10.1007/s42952-022-00189-x

[8] CHALONER, K. (1996). Elicitation of prior distributions. In *Bayesian Biostatistics* 141–156 Dekker, New York.

[9] CHALONER, K., CHURCH, T., LOUIS, T. A. and MATTS, J. P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *J. R. Stat. Soc., Ser. D, Stat.* **42** 341–353.

[10] CLEMEN, R. T. and REILLY, T. (1999). Correlations and copulas for decision and risk analysis. *Manage. Sci.* **45** 208–224.

[11] CONNOR, R. J. and MOSIMANN, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **64** 194–206. MR0240895

[12] CRISPINO, M. and ANTONIANO-VILLALOBOS, I. (2023). Informative priors for the consensus ranking in the Bayesian Mallows model. *Bayesian Anal.* **18** 391–414. MR4578058 https://doi.org/10.1214/22-ba1307

[13] DE SANTIS, F. (2007). Using historical data for Bayesian sample size determination. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **170** 95–113. MR2339100 https://doi.org/10.1111/j.1467-985X.2006.00438.x

[14] DEMARTA, S. and MCNEIL, A. J. (2005). The t copula and related copulas. *Int. Stat. Rev.* **73** 111–129.

[15] EBERLY, L. E. and CARLIN, B. P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat. Med.* **19** 2279–2294.

[16] ELFADALY, F. G. and GARTHWAITE, P. H. (2017). Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Stat. Comput.* **27** 449–467. MR3599683 https://doi.org/10.1007/s11222-016-9632-7

Luke Hagar is a Postdoctoral Scholar, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada (e-mail: lmhagar@uwaterloo.ca). Nathaniel T. Stevens is an Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada (e-mail: nstevens@uwaterloo.ca).

[17] Feroze, N. and Aslam, M. (2021). Comparison of improved class of priors for the analysis of the Burr type VII model under doubly censored samples. *Hacet. J. Math. Stat.* **50** 1509–1533. MR4331416 https://doi.org/10.15672/hujms.989479

[18] Freedman, L. and Spiegelhalter, D. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *J. R. Stat. Soc., Ser. D, Stat.* **32** 153–160.

[19] Fúquene, J., Steel, M. and Rossell, D. (2019). On choosing mixture components via non-local priors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 809–837. MR4025398 https://doi.org/10.1111/rssb.12333

[20] Gao, C., van der Vaart, A. W. and Zhou, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

[21] Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G. and Jenkinson, D. J. (2013). Prior distribution elicitation for generalized linear and piecewise-linear models. *J. Appl. Stat.* **40** 59–75. MR3042231 https://doi.org/10.1080/02664763.2012.734794

[22] Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical method for eliciting probability distributions. *J. Amer. Statist. Assoc.* **100** 680–701. MR2170464 https://doi.org/10.1198/016214505000000105

[23] Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94** 247–253. MR1689229 https://doi.org/10.2307/2669699

[24] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

[25] Gelman, A., Simpson, D. and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* **19** 555.

[26] Ghosal, S., Ghosh, J. K. and Samanta, T. (1995). On convergence of posterior distributions. *Ann. Statist.* **23** 2145–2152. MR1389869 https://doi.org/10.1214/aos/1034713651

[27] Gubbiotti, S. and De Santis, F. (2011). A Bayesian method for the choice of the sample size in equivalence trials. *Aust. N. Z. J. Stat.* **53** 443–460. MR2910028 https://doi.org/10.1111/j.1467-842X.2011.00643.x

[28] Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statist. Sci.* **20** 111–140. MR2183445 https://doi.org/10.1214/088342305000000098

[29] Gustafson, P. (2012). On the behaviour of Bayesian credible intervals in partially identified models. *Electron. J. Stat.* **6** 2107–2124. MR3020258 https://doi.org/10.1214/12-EJS741

[30] Hagar, L. and Stevens, N. T. (2024). Fast power curve approximation for posterior analyses. *Bayesian Anal.* https://doi.org/10.1214/24-BA1469

[31] Hagar, L. and Stevens, N. T. (2025). An economical approach to design posterior analyses. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2025.2476221

[32] Hagar, L. and Stevens, N. T. (2026). Supplement to "Posterior ramifications of prior dependence structures." https://doi.org/10.1214/25-STS981SUPPA, https://doi.org/10.1214/25-STS981SUPPB

[33] Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In *Distributions with Fixed Marginals and Related Topics (Seattle, WA*, 1993). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **28** 120–141. IMS, Hayward, CA. MR1485527 https://doi.org/10.1214/lnms/1215452614

[34] Joe, H. and Kurowicka, D. (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapore.

[35] Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T. and Feldman, B. M. (2010). Methods to elicit beliefs for Bayesian priors: A systematic review. *J. Clin. Epidemiol.* **63** 355–369.

[36] Jones, G. and Johnson, W. O. (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *Amer. Statist.* **68** 42–51. MR3303833 https://doi.org/10.1080/00031305.2013.868828

[37] Kadane, J. B. (1986). Progress toward a more ethical method for clinical trials. *J. Med. Philos.* **11** 385–404.

[38] Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75** 845–854. MR0600966

[39] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

[40] Kurowicka, D. and Cooke, R. M. (2005). Distribution-free continuous Bayesian belief nets. In *Modern Statistical and Mathematical Methods in Reliability*. *Ser. Qual. Reliab. Eng. Stat.* **10** 309–322. World Scientific, Singapore. MR2230715 https://doi.org/10.1142/9789812703378_0022

[41] Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1784901 https://doi.org/10.1007/978-1-4612-1166-2

[42] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

[43] Lemieux, C. (2009). Using quasi–Monte Carlo in practice. In *Monte Carlo and Quasi-Monte Carlo Sampling*. *Springer Series in Statistics* 1–46. Springer, New York.

[44] Libby, P. and Pober, J. S. (2001). Chronic rejection. *Immunity* **14** 387–397.

[45] Michimae, H. and Emura, T. (2022). Bayesian ridge estimators based on copula-based joint prior distributions for regression coefficients. *Comput. Statist.* **37** 2741–2769. MR4509625

[46] Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x

[47] Ning, S. and Shephard, N. (2018). A nonparametric Bayesian approach to copula estimation. *J. Stat. Comput. Simul.* **88** 1081–1105. MR3760400 https://doi.org/10.1080/00949655.2017.1421194

[48] O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, New York.

[49] Reimherr, M., Meng, X.-L. and Nicolae, D. L. (2021). Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 413–437. MR4294538 https://doi.org/10.1111/rssb.12414

[50] Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, Berlin.

[51] Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Stat.* **3** 395–402.

[52] Santos, J. D. and Costa, J. M. J. (2019). An algorithm for prior elicitation in dynamic Bayesian models for proportions with the logit link function. *Methodol. Comput. Appl. Probab.* **21** 169–183. MR3915437 https://doi.org/10.1007/s11009-018-9642-3

[53] Schillings, C., Sprungk, B. and Wacker, P. (2020). On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian

inverse problems. *Numer. Math.* **145** 915–971. MR4125981 https://doi.org/10.1007/s00211-020-01131-1

[54] SCHWEIZER, B. and SKLAR, A. (2011). *Probabilistic Metric Spaces*. Courier Corporation.

[55] SEO, J. I. and KIM, Y. (2022). Nonparametric prior elicitation for a binomial proportion. *Comm. Statist. Simulation Comput.* **51** 2809–2821. MR4434767 https://doi.org/10.1080/03610918.2019.1702210

[56] SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

[57] SRIVASTAVA, R., UPADHYAY, S. K. and SHUKLA, V. K. (2019). Subjective elicitation of hyperparameters of a conjugate Dirichlet prior and the corresponding Bayes analysis. *Comm. Statist. Theory Methods* **48** 4874–4887. MR3996108 https://doi.org/10.1080/03610926.2018.1500600

[58] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[59] VOS, R., VANAUDENAERDE, B., VERLEDEN, S. E., DE VLEESCHAUWER, S., WILLEMS-WIDYASTUTI, A., VAN RAEMDONCK, D., SCHOONIS, A., NAWROT, T., DUPONT, L. et al. (2011). A randomised controlled trial of azithromycin to prevent chronic rejection after lung transplantation. *Eur. Respir. J.* **37** 164–172.

[60] WILLIAMS, C. J., WILSON, K. J. and WILSON, N. (2021). A comparison of prior elicitation aggregation using the classical method and SHELF. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **184** 920–940. MR4305565 https://doi.org/10.1111/rssa.12691

[61] WILSON, K. J. (2018). Specification of informative prior distributions for multinomial models using vine copulas. *Bayesian Anal.* **13** 749–766. MR3807865 https://doi.org/10.1214/17-BA1068

[62] WILSON, K. J., ELFADALY, F. G., GARTHWAITE, P. H. and OAKLEY, J. E. (2021). Recent advances in the elicitation of uncertainty distributions from experts for multinomial probabilities. In *Expert Judgement in Risk and Decision Analysis. Internat. Ser. Oper. Res. Management Sci.* **293** 19–51. Springer, Cham. MR4238541 https://doi.org/10.1007/978-3-030-46474-5_2

[63] WINKLER, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc.* **62** 776–800. MR0220368

[64] WINKLER, R. L. (1971). Probabilistic prediction: Some experimental results. *J. Amer. Statist. Assoc.* **66** 675–685.

[65] WONG, W. H. and MA, L. (2010). Optional Pólya tree and Bayesian inference. *Ann. Statist.* **38** 1433–1459. MR2662348 https://doi.org/10.1214/09-AOS755

[66] WU, J., WANG, X. and WALKER, S. G. (2015). Bayesian nonparametric estimation of a copula. *J. Stat. Comput. Simul.* **85** 103–116. MR3266006 https://doi.org/10.1080/00949655.2013.806508

[67] YE, K., HAN, Z., DUAN, Y. and BAI, T. (2022). Normalized power prior Bayesian analysis. *J. Statist. Plann. Inference* **216** 29–50. MR4269624 https://doi.org/10.1016/j.jspi.2021.05.005

Statistical Science
2026, Vol. 41, No. 3, 524–560
https://doi.org/10.1214/25-STS984

Evaluating Machine Learning Models in Nonstandard Settings: An Overview and New Findings

Roman Hornung, Malte Nalenz, Lennart Schneider, Andreas Bender, Ludwig Bothmann, Florian Dumpert, Bernd Bischl, Thomas Augustin and Anne-Laure Boulesteix

Abstract. Estimating the generalization error (GE) of machine learning models is fundamental, with resampling methods being the most common approach. However, in nonstandard settings, particularly those where observations are not independently and identically distributed, resampling using simple random data divisions may lead to biased GE estimates. This paper strives to present well-grounded guidelines for GE estimation in various such nonstandard settings: clustered data, spatial data, unequal sampling probabilities, concept drift and hierarchically structured outcomes. Our overview combines well-established methodologies with other existing methods that, to our knowledge, have not been frequently considered in these particular settings. A unifying principle among these techniques is that the test data used in each iteration of the resampling procedure should reflect the new observations to which the model will be applied, while the training data should be representative of the entire data set used to obtain the final model. Beyond providing an overview, we address literature gaps by conducting simulation studies and a real data study. These studies assess the necessity of using GE-estimation methods tailored to the respective setting. Our findings corroborate the concern that standard resampling methods often yield biased GE estimates in nonstandard settings, underscoring the importance of tailored GE estimation.

Key words and phrases: Generalization error, overoptimism, resampling, cross-validation, machine learning, spatial data, clustered data, unequal sampling probabilities, concept drift, hierarchical classification.

Roman Hornung is Senior Research Fellow, LMU Munich, Faculty of Medicine, Institute for Medical Information Processing, Biometry and Epidemiology, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: hornung@ibe.med.uni-muenchen.de). Malte Nalenz was Senior Research Fellow at the LMU Munich, Department of Statistics, Munich, Germany, while doing this work (e-mail: malte.nlz@gmail.com). Lennart Schneider is Research Fellow, LMU Munich, Department of Statistics, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: lennart.schneider@stat.uni-muenchen.de). Andreas Bender is Senior Research Fellow, LMU Munich, Department of Statistics, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: andreas.bender@stat.uni-muenchen.de). Ludwig Bothmann is Senior Research Fellow, LMU Munich, Department of Statistics, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: ludwig.bothmann@stat.uni-muenchen.de). Florian Dumpert is head of the Section for Artificial Intelligence and Big Data, Federal Statistical Office of Germany, Wiesbaden, Germany (e-mail: florian.dumpert@destatis.de). Bernd Bischl is Professor, LMU Munich, Department of Statistics, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: bernd.bischl@stat.uni-muenchen.de). Thomas Augustin is Professor, LMU Munich, Department of Statistics, Munich, Germany (e-mail: thomas.augustin@stat.uni-muenchen.de). Anne-Laure Boulesteix is Professor, LMU Munich, Faculty of Medicine, Institute for Medical Information Processing, Biometry and Epidemiology, Munich Center for Machine Learning (MCML), Munich, Germany (e-mail: boulesteix@ibe.med.uni-muenchen.de).

REFERENCES

ALY, R., REMUS, S. and BIEMANN, C. (2019). Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*: *Student Research Workshop* 323–330.

AUSTERN, M. and ZHOU, W. (2020). Asymptotics of cross-validation. Preprint. Available at arXiv:2001.11111.

BARREÑADA, L., DHIMAN, P., TIMMERMAN, D., BOULESTEIX, A. L. and VAN CALSTER, B. (2024). Understanding overfitting in random forest for probability estimation: A visualization and simulation study. *Diagn*. *Progn*. *Res* **8** 14.

BATES, S., HASTIE, T. and TIBSHIRANI, R. (2024). Cross-validation: What does it estimate and how well does it do it? *J*. *Amer*. *Statist*. *Assoc*. **119** 1434–1445. MR4766000 https://doi.org/10.1080/01621459.2023.2197686

BAYLE, P., BAYLE, A., JANSON, L. and MACKEY, L. (2020). Cross-validation confidence intervals for test error. *Adv*. *Neur*. *In*. **33** 16339–16350.

BENGIO, Y. and GRANDVALET, Y. (2003/04). No unbiased estimator of the variance of K-fold cross-validation. *J*. *Mach*. *Learn*. *Res*. **5** 1089–1105. MR2248010

BENNETT, P. N. and NGUYEN, N. (2009). Refined experts: Improving classification in large taxonomies. In *Proceedings of the* 32*nd International ACM SIGIR Conference on Research and Development in Information Retrieval* 11–18.

BERGMEIR, C., HYNDMAN, R. J. and KOO, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput*. *Statist*. *Data Anal*. **120** 70–83. MR3742209 https://doi.org/10.1016/j.csda.2017.11.003

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

BIFET, A. and GAVALDA, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the* 2007 *SIAM International Conference on Data Mining* 443–448.

BISCHL, B., BINDER, M., LANG, M., PIELOK, T., RICHTER, J., COORS, S., THOMAS, J., ULLMANN, T., BECKER, M. et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Min*. *Knowl*. **13** e1484.

BISCHL, B., MERSMANN, O., TRAUTMANN, H. and WEIHS, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evol*. *Comput*. **20** 249–275.

BISCHL, B., SONABEND, R., KOTTHOFF, L. and LANG, M., eds. (2024). In *Applied Machine Learning Using mlr*3 *in R* 3. CRC Press, Boca Raton.

BLOCKEEL, H., BRUYNOOGHE, M., DŽEROSKI, S., RAMON, J. and STRUYF, J. (2002). Hierarchical multi-classification. In *Workshop Notes of the KDD'*02 *Workshop on Multi-Relational Data Mining* 21–35.

BOULESTEIX, A. L., STROBL, C., AUGUSTIN, T. and DAUMER, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Inform*. **6** 77–97.

BRAGA-NETO, U. M. and DOUGHERTY, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20** 374–380.

BREIDT, F. J. and OPSOMER, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statist*. *Sci*. **32** 190–205. MR3648955 https://doi.org/10.1214/16-STS589

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

BRENNING, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Nat*. *Hazard Earth Sys*. **5** 853–862.

BRENNING, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In 2012 *IEEE International Geoscience and Remote Sensing Symposium* 5372–5375.

BRENNING, A. and LAUSEN, B. (2008). Estimating error rates in the classification of paired organs. *Stat*. *Med*. **27** 4515–4531. MR2528528 https://doi.org/10.1002/sim.3310

CAI, L. and HOFMANN, T. (2007). Exploiting known taxonomies in learning overlapping concepts. In *Proceedings of the* 20*th International Joint Conference on Artifical Intelligence* 714–719.

CESA-BIANCHI, N., GENTILE, C. and ZANIBONI, L. (2004). Incremental algorithms for hierarchical classification. In *Advances in Neural Information Processing Systems* (L. K. Saul, Y. Weiss and L. Bottou, eds.) **17** 233–240. MIT Press, Cambridge, MA.

CHARUVAKA, A. and RANGWALA, H. (2015). HierCost: Improving large scale hierarchical classification with cost sensitive learning. In *Proceedings of the* 2015 *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 675–690.

COSTA, E. P., LORENA, A. C., CARVALHO, A. C. P. L. F. and FREITAS, A. A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for Machine Learning II*: *Papers from the AAAI*-2007 *Workshop* (C. Drummond, W. Elazmeh, N. Japkowicz and S. A. Macskassy, eds.) 1–6. AAAI Press, Menlo Park.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151

DAGDOUG, M., GOGA, C. and HAZIZA, D. (2023). Model-assisted estimation through random forests in finite population sampling. *J*. *Amer*. *Statist*. *Assoc*. **118** 1234–1251. MR4595490 https://doi.org/10.1080/01621459.2021.1987250

DAWID, A. P. (1984). Statistical theory. The prequential approach. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

DOBBIN, K. K. and SIMON, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med*. *Genom*. **4** 1–8.

DUCANGE, P., MARCELLONI, F. and PECORI, R. (2021). Fuzzy Hoeffding decision tree for data stream classification. *Int*. *J*. *Comput*. *Intell*. *Syst*. **14** 946–964.

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *J*. *Amer*. *Statist*. *Assoc*. **92** 548–560. MR1467848 https://doi.org/10.2307/2965703

ELLENBACH, N., BOULESTEIX, A.-L., BISCHL, B., UNGER, K. and HORNUNG, R. (2021). Improved outcome prediction across data sources through robust parameter tuning. *J*. *Classification* **38** 212–231. MR4290431 https://doi.org/10.1007/s00357-020-09368-z

EUROPEAN COMMISSION (2023). The impact of demographic change—in a changing environment. Technical Report.

FREITAS, A. A. and DE CARVALHO, A. C. (2007). A tutorial on hierarchical classification with applications in bioinformatics. In *Research and Trends in Data Mining Technologies and Applications* (D. Taniar, ed.) 175–208 7. Idea Group.

FRIAS-BLANCO, I., DEL CAMPO-AVILA, J., RAMOS-JIMÉNEZ, G., CARVALHO, A. C., ORTIZ-DÍAZ, A. and MORALES-BUENO, R. (2016). Online adaptive decision trees based on concentration inequalities. *Knowl*.-*Based Syst*. **104** 179–194.

GAMA, J., MEDAS, P., CASTILLO, G. and RODRIGUES, P. (2004). Learning with drift detection. In *Proceedings of the* 17*th Brazilian Symposium on Artificial Intelligence* 286–295.

GAMA, J., ROCHA, R. and MEDAS, P. (2003). Accurate decision trees for mining high-speed data streams. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 523–528.

GAMA, J., SEBASTIÃO, R. and RODRIGUES, P. P. (2013). On evaluating stream learning algorithms. *Mach*. *Learn*. **90** 317–346. MR3021877 https://doi.org/10.1007/s10994-012-5320-9

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M. and BOUCHACHIA, A. (2014). A survey on concept drift adaptation. *ACM Comput*. *Surv*. **46** 1–37.

GARGIULO, F., SILVESTRI, S., CIAMPI, M. and DE PIETRO, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Appl*. *Soft Comput*. **79** 125–138.

GHOLAMIANGONABADI, D., KISELOV, N. and GROLINGER, K. (2020). Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *IEEE Access* **8** 133982–133994.

GONÇALVES JR., P. M., DE CARVALHO SANTOS, S. G., BARROS, R. S. and VIEIRA, D. C. (2014). A comparative study on concept drift detectors. *Expert Syst*. *Appl*. **41** 8144–8156.

GOPAL, S. and YANG, Y. (2013). Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 257–265.

HÁJEK, J. (1971). Comment on "An essay on the logical foundations of survey sampling". In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 236. Holt, Rinehart and Winston, New York.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HAWKINS, D. M., BASAK, S. C. and MILLS, D. (2003). Assessing model fit by cross-validation. *J*. *Chem*. *Inf*. *Comput*. *Sci*. **43** 579–586.

HAZIZA, D. and BEAUMONT, J.-F. (2017). Construction of weights in surveys: A review. *Statist*. *Sci*. **32** 206–226. MR3648956 https://doi.org/10.1214/16-STS608

HEIKKINEN, R. K., MARMION, M. and LUOTO, M. (2012). Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography* **35** 276–288.

HOLBROOK, A., LUMLEY, T. and GILLEN, D. (2020). Estimating prediction error for complex samples. *Canad*. *J*. *Statist*. **48** 204–221. MR4095260 https://doi.org/10.1002/cjs.11527

HORNUNG, R., NALENZ, M., SCHNEIDER, L., BENDER, A., BOTHMANN, L., BISCHL, B., AUGUSTIN, T. and BOULESTEIX, A.-L. (2026). Supplement to "Evaluating machine learning models in nonstandard settings: an overview and new findings." https://doi.org/10.1214/25-STS984SUPPA, https://doi.org/10.1214/25-STS984SUPPB

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J*. *Amer*. *Statist*. *Assoc*. **47** 663–685. MR0053460

HUANG, G.-B., ZHU, Q.-Y. and SIEW, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing* **70** 489–501.

HUANG, W., CHEN, E., LIU, Q., CHEN, Y., HUANG, Z., LIU, Y., ZHAO, Z., ZHANG, D. and WANG, S. (2019). Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management* 1051–1060.

HULTEN, G., SPENCER, L. and DOMINGOS, P. (2001). Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 97–106.

HYNDMAN, R. J. and ATHANASOPOULOS, G. (2018). *Forecasting*: *Principles and Practice*. OTexts, Melbourne.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (2022). Climate Change 2022: Impacts, Adaptation and Vulnerability. Technical Report.

INTERNATIONAL LABOUR ORGANIZATION (2012). International standard classification of occupations: ISCO-08. Technical Report.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning*. *Springer Texts in Statistics* **103**. Springer, New York. MR3100153 https://doi.org/10.1007/978-1-4614-7138-7

JAPKOWICZ, N. and SHAH, M. (2011). *Evaluating Learning Algorithms*: *A Classification Perspective*. Cambridge Univ. Press, Cambridge.

KIRITCHENKO, S., MATWIN, S. and FAMILI, A. F. (2005). Functional annotation of genes using hierarchical text categorization. In *Proceedings of the BioLINK SIG*: *Linking Literature*, *Information and Knowledge for Biology* (*Held at ISMB*-05).

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence* 1137–1143.

KOSMOPOULOS, A., PARTALAS, I., GAUSSIER, E., PALIOURAS, G. and ANDROUTSOPOULOS, I. (2015). Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Min*. *Knowl*. *Discov*. **29** 820–865. MR3334322 https://doi.org/10.1007/s10618-014-0382-x

KOWSARI, K., BROWN, D. E., HEIDARYSAFA, M., JAFARI MEIMANDI, K., GERBER, M. S. and BARNES, L. E. (2017). HDLTex: Hierarchical deep learning for text classification. In *Proceedings of the* 16*th IEEE International Conference on Machine Learning and Applications* (*ICMLA*) 364–371.

KUNJAN, S., GRUMMETT, T. S., POPE, K. J., POWERS, D. M., FITZGIBBON, S. P., BASTIAMPILLAI, T., BATTERSBY, M. and LEWIS, T. W. (2021). The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis. In *Proceedings of the* 14*th International Conference on Brain Informatics*, *Virtual Event* 558–567.

KÜNZEL, S. H., MÖLLER, P. T., LINDNER, M., GOERDT, L., NADAL, J., SCHMID, M., SCHMITZ-VALCKENBERG, S., HOLZ, F. G., FLECKENSTEIN, M. et al. (2020). Determinants of quality of life in geographic atrophy secondary to age-related macular degeneration. *Investig*. *Ophthalmol*. *Vis*. *Sci*. **61** 63–63.

LE REST, K., PINAUD, D., MONESTIEZ, P., CHADOEUF, J. and BRETAGNOLLE, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Glob*. *Ecol*. *Biogeogr*. **23** 811–820.

LU, J., LIU, A., DONG, F., GU, F., GAMA, J. and ZHANG, G. (2018). Learning under concept drift: A review. *IEEE Trans*. *Knowl*. *Data Eng*. **31** 2346–2363.

MAO, Y., TIAN, J., HAN, J. and REN, X. (2019). Hierarchical text classification with reinforced label assignment. In *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*) 445–455.

MCCALLUM, A., ROSENFELD, R., MITCHELL, T. M. and NG, A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning* 359–367.

MEERTENS, Q., DIKS, C., VAN DEN HERIK, H. and TAKES, F. (2022). Improving the output quality of official statistics based on machine learning algorithms. *J*. *Off*. *Stat*. **38** 485–508.

MINKU, L. L., WHITE, A. P. and YAO, X. (2009). The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Trans*. *Knowl*. *Data Eng*. **22** 730–742.

MOLINARO, A. M., SIMON, R. and PFEIFFER, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **21** 3301–3307.

MORITZ, S., DUMPERT, F., JUNG, C., BARTZ-BEIELSTEIN, T. and BARTZ, E. (2024). Practical applications of online machine learning. In *Online Machine Learning*: *A Practical Guide with Examples in Python* 71–96. Springer Nature Singapore, Singapore.

NALENZ, M., RODEMANN, J. and AUGUSTIN, T. (2024). Learning de-biased regression trees and forests from complex samples. *Mach*. *Learn*. **113** 3379–3398. MR4745078 https://doi.org/10.1007/s10994-023-06439-1

PENG, H., LI, J., HE, Y., LIU, Y., BAO, M., WANG, L., SONG, Y. and YANG, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the* 2018 *World Wide Web Conference* 1063–1072.

PFAU, M., VON DER EMDE, L., DYSLI, C., MÖLLER, P. T., THIELE, S., LINDNER, M., SCHMID, M., RUBIN, D. L., FLECKENSTEIN, M. et al. (2020). Determinants of cone and rod functions in geographic atrophy: AI-based structure-function correlation. *Am*. *J*. *Ophthalmol*. **217** 162–173.

RASCHKA, S. (2020). Model evaluation, model selection, and algorithm selection in machine learning. Preprint. Available at arXiv:1811.12808.

ROBERTS, D. R., BAHN, V., CIUTI, S., BOYCE, M. S., ELITH, J., GUILLERA-ARROITA, G., HAUENSTEIN, S., LAHOZ-MONFORT, J. J., SCHRÖDER, B. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40** 913–929.

RUSS, G. and BRENNING, A. (2010). Data mining in precision agriculture: Management of spatial information. In *Proceedings of the* 13*th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* 350–359.

RUTKOWSKI, L., JAWORSKI, M., PIETRUCZUK, L. and DUDA, P. (2015). A new method for data stream mining based on the misclassification error. *IEEE Trans*. *Neural Netw*. *Learn*. *Syst*. **26** 1048–1059. MR3454262 https://doi.org/10.1109/TNNLS.2014.2333557

SAEB, S., LONINI, L., JAYARAMAN, A., MOHR, D. C. and KORDING, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience* **6** gix019.

SAIDANI, Y., DUMPERT, F., BORGS, C., BRAND, A., NICKL, A., RITTMANN, A., ROHDE, J., SALWICZEK, C., STORFINGER, N. et al. (2023). Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik. *AStA Wirtsch*. *Sozialstat*. *Arch*. **17** 253–303.

SÄRNDAL, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67** 639–650. MR0601102 https://doi.org/10.2307/2335134

SCHRATZ, P. and BECKER, M. (2022). mlr3spatiotempcv: Spatiotemporal resampling methods for 'mlr3'. R package version 2.0.3.

SCHRATZ, P., BECKER, M., LANG, M. and BRENNING, A. (2021). Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R. Preprint. Available at arXiv:2110.12674.

SCHRATZ, P., MUENCHOW, J., ITURRITXA, E., RICHTER, J. and BRENNING, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol*. *Model*. **406** 109–120.

SCHREUDER, H. T., GREGOIRE, T. G. and WEYER, J. P. (2001). For what applications can probability and non-probability sampling be used? *Environ*. *Monit*. *Assess*. **66** 281–291.

SCHULZ-KÜMPEL, H., FISCHER, S., NAGLER, T., BOULESTEIX, A.-L., BISCHL, B. and HORNUNG, R. (2024). Constructing Confidence Intervals for 'the' Generalization Error—a Comprehensive Benchmark Study. Preprint. Available at arXiv:2409.18836.

SIMON, R. (2007). Resampling strategies for model assessment and selection. In *Fundamentals of Data Mining in Genomics and Proteomics* (W. Dubitzky, M. Granzow and D. Berrar, eds.) 173–186. Springer, Boston, MA.

SKINNER, C. and WAKEFIELD, J. (2017). Introduction to the design and analysis of complex survey data. *Statist*. *Sci*. **32** 165–175. MR3648953 https://doi.org/10.1214/17-STS614

SOKOLOVA, M. and LAPALME, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf*. *Process*. *Manag*. **45** 427–437.

STEYERBERG, E. W. (2019). *Clinical Prediction Models*: *A Practical Approach to Development*, *Validation*, *and Updating*, 2nd ed. Springer, New York.

SUN, A. and LIM, E.-P. (2001). Hierarchical text classification and evaluation. In *Proceedings of the* 2001 *IEEE International Conference on Data Mining* 521–528.

TOTH, D. and ELTINGE, J. L. (2011). Building consistent regression trees from complex sample data. *J*. *Amer*. *Statist*. *Assoc*. **106** 1626–1636. MR2896862 https://doi.org/10.1198/jasa.2011.tm10383

TOUGUI, I., JILBAB, A. and EL MHAMDI, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc*. *Inform*. *Res*. **27** 189–199.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (2011). The Impact of Globalization on National Accounts. Technical Report.

UNITED NATIONS—DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS (2022). The Handbook on Management and Organization of National Statistical Systems. Technical Report.

VACH, W. (2013). *Regression Models as a Tool in Medical Research*. CRC Press, Boca Raton, FL. MR3013875

VALAVI, R., ELITH, J., LAHOZ-MONFORT, J. J. and GUILLERA-ARROITA, G. (2018). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol*. *Evol*. **10**.

VALLIANT, R., DEVER, J. A. and KREUTER, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. *Statistics for Social and Behavioral Sciences*. Springer, Cham. MR3887634 https://doi.org/10.1007/978-3-319-93632-1

WANG, K., ZHOU, S. and LIEW, S. C. (1999). Building hierarchical classifiers using class proximity. In *Proceedings of the* 25*th International Conference on Very Large Data Bases* 363–374.

WEBB, G. I., HYDE, R., CAO, H., NGUYEN, H. L. and PETITJEAN, F. (2016). Characterizing concept drift. *Data Min*. *Knowl*. *Discov*. **30** 964–994. MR3513248 https://doi.org/10.1007/s10618-015-0448-4

WEHRMANN, J., CERRI, R. and BARROS, R. (2018). Hierarchical multi-label classification networks. In *Proceedings of the* 35*th International Conference on Machine Learning* 5075–5084.

WENGER, S. J. and OLDEN, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods Ecol*. *Evol*. **3** 260–267.

WORLD BANK GROUP (2019). *The Changing Nature of Work*. *World Development Report*. World Bank.

WU, C., TYGERT, M. and LECUN, Y. (2019). A hierarchical loss and its problems when classifying non-hierarchically. *PLoS ONE* **14** e0226222.

XU, S. and WANG, J. (2017). Dynamic extreme learning machine for data stream classification. *Neurocomputing* **238** 433–449.

XUE, G.-R., XING, D., YANG, Q. and YU, Y. (2008). Deep classification in large-scale text hierarchies. In *Proceedings of the* 31*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 619–626.

YUNG, W., TAM, S.-M., BUELENS, B., CHIPMAN, H., DUMPERT, F., ASCARI, G., ROCCI, F., BURGER, J. and CHOI, I. K. (2022). A quality framework for statistical algorithms. *Stat. J. IAOS* **38** 291–308.

ZHANG, C., ZHANG, Y., SHI, X., ALMPANIDIS, G., FAN, G. and SHEN, X. (2019). On incremental learning for gradient boosting decision trees. *Neural Process. Lett.* **50** 957–987.

ZHANG, Y., CHU, G., LI, P., HU, X. and WU, X. (2017). Three-layer concept drifting detection in text data streams. *Neurocomputing* **260** 393–403.

ZHOU, D., XIAO, L. and WU, M. (2011). Hierarchical classification via orthogonal transfer. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*) 801–808.

ŽLIOBAITĖ, I. (2014). Controlled permutations for testing adaptive learning models. *Knowl. Inf. Syst.* **39** 565–578.

Statistical Science
2026, Vol. 41, No. 3, 561–581
https://doi.org/10.1214/25-STS985

A Review of Off-Policy Evaluation in Reinforcement Learning

Masatoshi Uehara, Chengchun Shi and Nathan Kallus

Abstract. Reinforcement learning (RL) is one of the most vibrant research frontiers in machine learning and has been recently applied to solve a number of challenging problems. In this paper, we primarily focus on off-policy evaluation (OPE), one of the most fundamental topics in RL. In recent years, a number of OPE methods have been developed in the statistics and computer science literature. We provide a discussion on the efficiency bound of OPE, some of the existing state-of-the-art OPE methods, their statistical properties and some other related research directions that are currently actively explored.

Key words and phrases: Off-policy evaluation, semiparametric methods, causal inference, dynamic treatment regime, offline reinforcement learning, contextual bandits.

REFERENCES

[1] AGARWAL, A., BASU, S., SCHNABEL, T. and JOACHIMS, T. (2017). Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the* 23rd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 687–696.

[2] AGARWAL, A., JIANG, N., KAKADE, S. M. and SUN, W. (2019). Reinforcement learning: Theory and algorithms. CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep.

[3] AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420 https://doi.org/10.1111/1468-0262.00470

[4] AI, C. and CHEN, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *J. Econometrics* **170** 442–457. MR2970325 https://doi.org/10.1016/j.jeconom.2012.05.015

[5] ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2006). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *Learning Theory*. *Lecture Notes in Computer Science* **4005** 574–588. Springer, Berlin. MR2280632 https://doi.org/10.1007/11776420_42

[6] ATHEY, S. and WAGER, S. (2017). Efficient Policy Learning. arXiv preprint. Available at arXiv:1702.02896.

[7] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

[8] BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In 2016 *IEEE International Conference on Data*. *Science and Advanced Analytics* (*DSAA*) **2016** 689–696. IEEE.

[9] BENNETT, A. and KALLUS, N. (2021). Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes.

[10] BENNETT, A., KALLUS, N., LI, L. and MOUSAVI, A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics* 1999–2007. PMLR.

[11] BERTSEKAS, D. P. (2012). *Dynamic Programming and Optimal Control*. *Athena Scientific Optimization and Computation Series*, 4th edn. Athena Scientific, Belmont, MA.

[12] BERTSEKAS, D. P. and YU, H. (2009). Projected equation methods for approximate solution of large linear systems. *J. Comput. Appl. Math.* **227** 27–50. MR2512758 https://doi.org/10.1016/j.cam.2008.07.037

[13] BHATTACHARYA, R., NABI, R. and SHPITSER, I. (2022). Semiparametric inference for causal effects in graphical models with hidden variables. *J. Mach. Learn. Res.* **23** Paper No. [295], 76. MR4577734

[14] BIBAUT, A., MALENICA, I., VLASSIS, N. and VAN DER LAAN, M. (2019). More efficient off-policy evaluation through regularized targeted learning. In *Proceedings of the* 36th *International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **97** 654–663.

[15] BIBAUT, A., PETERSEN, M., VLASSIS, N., DIMAKOPOULOU, M. and VAN DER LAAN, M. (2021). Sequential causal inference in a single world of connected units. arXiv preprint. Available at arXiv:2101.07380.

[16] BIBAUT, A. and VAN DER LAAN, M. J. (2017). Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. arXiv preprint. Available at arXiv:1706.07408.

Masatoshi Uehara works for EvolutionaryScale, New York, New York, USA (e-mail: ueharamasatoshi136@gmail.com). Chengchun Shi is an Associate Professor in the Department of Statistics at London School of Economics and Political Science, London, UK (e-mail: c.shi7@lse.ac.uk). Nathan Kallus is an Associate Professor of Operations Research and Information Engineering, Cornell Tech, New York, New York 10044, USA (e-mail: kallus@cornell.edu).

[17] BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274
[18] CAO, W., TSIATIS, A. A. and DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768 https://doi.org/10.1093/biomet/asp033
[19] CHAKRABORTY, B., LABER, E. B. and ZHAO, Y.-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials* **11** 408–417.
[20] CHAMBERLAIN, G. (1992). Comment: Sequential moment restrictions in panel data. *J. Bus. Econom. Statist.* **10** 20–26.
[21] CHAPELLE, O. and LI, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems* **24** 2249–2257.
[22] CHEN, E. Y., SONG, R. and JORDAN, M. I. (2024). Reinforcement learning in latent heterogeneous environments. *J. Amer. Statist. Assoc.* **119** 3113–3126. MR4833941 https://doi.org/10.1080/01621459.2024.2308317
[23] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the* 36*th International Conference on Machine Learning* **97** 1042–1051.
[24] CHEN, X. and QI, Z. (2022). On well-posedness and minimax optimal rates of nonparametric Q-function estimation in off-policy evaluation. arXiv preprint. Available at arXiv:2201.06169.
[25] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097
[26] CHERNOZHUKOV, V., NEWEY, W., SINGH, R. and SYRGKANIS, V. (2020). Adversarial estimation of Riesz representers. arXiv preprint. Available at arXiv:2101.00009.
[27] CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *Econom. J.* **25** 576–601. MR4565439 https://doi.org/10.1093/ectj/utac002
[28] CLIFTON, J. and LABER, E. (2020). *Q*-learning: Theory and applications. *Annu. Rev. Stat. Appl.* **7** 279–301. MR4104194 https://doi.org/10.1146/annurev-statistics-031219-041220
[29] DAI, B., NACHUM, O., CHOW, Y., LI, L., SZEPESVÁRI, C. and SCHUURMANS, D. (2020). Coindice: Off-policy confidence interval estimation. arXiv preprint. Available at arXiv:2010.11652.
[30] DANN, C., NEUMANN, G. and PETERS, J. (2014). Policy evaluation with temporal differences: A survey and comparison. *J. Mach. Learn. Res.* **15** 809–883. MR3195332
[31] DÍAZ, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* **21** 353–358. MR4132549 https://doi.org/10.1093/biostatistics/kxz042
[32] DIKKALA, N., LEWIS, G., MACKEY, L. and SYRGKANIS, V. (2020). arXiv preprint. Available at arXiv:2006.07201.
[33] DU, S. S., KAKADE, S. M., WANG, R. and YANG, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? arXiv preprint. Available at arXiv:1910.03016.
[34] DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* 2701–2709. PMLR.
[35] DUAN, Y., JIN, C. and LI, Z. (2021). Risk Bounds and Rademacher Complexity in Batch Reinforcement Learning. arXiv preprint. Available at arXiv:2103.13883.
[36] DUDÍK, M., ERHAN, D., LANGFORD, J. and LI, L. (2014). Doubly robust policy evaluation and optimization. *Statist. Sci.* **29** 485–511. MR3300356 https://doi.org/10.1214/14-STS500
[37] ERNST, D., GEURTS, P. and WEHENKEL, L. (2005). Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.* **6** 503–556. MR2249830
[38] ERTEFAIE, A. and STRAWDERMAN, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* **105** 963–977. MR3877877 https://doi.org/10.1093/biomet/asy043
[39] FAN, C., LU, W., SONG, R. and ZHOU, Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1565–1582. MR3731676 https://doi.org/10.1111/rssb.12216
[40] FAN, J., WANG, Z., XIE, Y. and YANG, Z. (2020). A theoretical analysis of deep Q-learning. In *Proceedings of the* 2*nd Conference on Learning for Dynamics and Control. Proceedings of Machine Learning Research* **120** 486–489.
[41] FARAHMAND, A., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.* **17** Paper No. 139, 66. MR3555030
[42] FARAJTABAR, M., CHOW, Y. and GHAVAMZADEH, M. (2018). More robust doubly robust off-policy evaluation. In *Proceedings of the* 35*th International Conference on Machine Learning* 1447–1456.
[43] FENG, Y., LI, L. and LIU, Q. (2019). A kernel loss for solving the Bellman equation. In *Advances in Neural Information Processing Systems* **32** 15430–15441.
[44] FOSTER, D. J., KRISHNAMURTHY, A., SIMCHI-LEVI, D. and XU, Y. (2021). Offline reinforcement learning: Fundamental barriers for value function approximation. arXiv preprint. Available at arXiv:2111.10919.
[45] FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258
[46] FU, Z., QI, Z., WANG, Z., YANG, Z., XU, Y. and KOSOROK, M. R. (2022). Offline reinforcement learning with instrumental variables in confounded Markov decision processes. arXiv preprint. Available at arXiv:2209.08666.
[47] FULCHER, I. R., SHPITSER, I., MAREALLE, S. and TCHETGEN TCHETGEN, E. J. (2020). Robust inference on population indirect causal effects: The generalized front door criterion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 199–214. MR4060982
[48] FUTOMA, J., HUGHES, M. C. and DOSHI-VELEZ, F. (2020). Popcorn: Partially observed prediction constrained reinforcement learning. arXiv preprint. Available at arXiv:2001.04032.
[49] GOTTESMAN, O., JOHANSSON, F., KOMOROWSKI, M., FAISAL, A., SONTAG, D., DOSHI-VELEZ, F. and CELI, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nat. Med.* **25** 16–18.
[50] HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560
[51] HAO, B., DUAN, Y., LATTIMORE, T., SZEPESVÁRI, C. and WANG, M. (2021). Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning* 4063–4073. PMLR.
[52] HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998). Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* **65** 261–294. MR1623713 https://doi.org/10.1111/1467-937X.00044
[53] HERNAN, M. A. and ROBINS, J. M. (2019). *Causal Inference*. CRC Press, Boca Raton, FL.

[54] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

[55] HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

[56] HU, X., QIAN, M., CHENG, B. and CHEUNG, Y. K. (2021). Personalized policy learning using longitudinal mobile health data. *J. Amer. Statist. Assoc.* **116** 410–420. MR4227703 https://doi.org/10.1080/01621459.2020.1785476

[57] HU, Y. and WAGER, S. (2023). Off-policy evaluation in partially observed Markov decision processes under sequential ignorability. *Ann. Statist.* **51** 1561–1585. MR4658568 https://doi.org/10.1214/23-aos2287

[58] HUANG, A. and JIANG, N. (2022). Beyond the return: Off-policy function estimation under user-specified error-measuring distributions. In *Neurips*.

[59] HUBER, M. (2019). An introduction to flexible methods for policy evaluation. arXiv preprint. Available at arXiv:1910.00641.

[60] IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941 https://doi.org/10.1111/rssb.12027

[61] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[62] JIANG, H., DAI, B., YANG, M., ZHAO, T. and WEI, W. (2021). Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. arXiv preprint. Available at arXiv:2102.10242.

[63] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume* 652–661.

[64] JIANG, R., LU, W., SONG, R. and DAVIDIAN, M. (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1165–1185. MR3689313 https://doi.org/10.1111/rssb.12201

[65] JIN, Y., YANG, Z. and WANG, Z. (2020). Is pessimism provably efficient for offline RL? arXiv preprint. Available at arXiv:2012.15085.

[66] JUNG, Y., TIAN, J. and BAREINBOIM, E. (2020). Learning causal effects via weighted empirical risk minimization. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 12697–12709. Curran Associates, Red Hook.

[67] KALLUS, N. and UEHARA, M. (2019). Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems* **32** 3320–3329.

[68] KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21** Paper No. 167, 63. MR4209453

[69] KALLUS, N. and UEHARA, M. (2020). Doubly robust off-policy value and gradient estimation for deterministic policies. arXiv preprint. Available at arXiv:2006.03900.

[70] KALLUS, N. and UEHARA, M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.* **70** 3282–3302. MR4538517

[71] KALLUS, N. and UEHARA, M. (2024). Efficient evaluation of natural stochastic policies in off-line reinforcement learning. *Biometrika* **111** 51–69. MR4704558 https://doi.org/10.1093/biomet/asad059

[72] KALLUS, N. and ZHOU, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems* **33** 22293–22304. Curran Associates, Red Hook.

[73] KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

[74] KENNEDY, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *J. Amer. Statist. Assoc.* **114** 645–656. MR3963169 https://doi.org/10.1080/01621459.2017.1422737

[75] KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint. Available at arXiv:2203.06469.

[76] KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. MR3689316 https://doi.org/10.1111/rssb.12212

[77] KIDAMBI, R., RAJESWARAN, A., NETRAPALLI, P. and JOACHIMS, T. (2020). MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems* **33** 21810–21823. Curran Associates, Red Hook.

[78] KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* **86** 591–616. MR3783340 https://doi.org/10.3982/ECTA13288

[79] KOBER, J., BAGNELL, J. A. and PETERS, J. (2013). Reinforcement learning in robotics: A survey. *Int. J. Robot. Res.* **32** 1238–1274.

[80] KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

[81] KOSOROK, M. R. and MOODIE, E. E. M., eds. (2015). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Society for Industrial and Applied Mathematics, USA.

[82] KUZBORSKIJ, I., VERNADE, C., GYORGY, A. and SZEPESVÁRI, C. (2021). Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics* 640–648. PMLR.

[83] LAGOUDAKIS, M. G. and PARR, R. (2004). Least-squares policy iteration. *J. Mach. Learn. Res.* **4** 1107–1149. MR2125347 https://doi.org/10.1162/1532443041827907

[84] LE, H., VOLOSHIN, C. and YUE, Y. (2019). Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning* 3703–3712.

[85] LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

[86] LEVINE, S., KUMAR, A., TUCKER, G. and FU, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint. Available at arXiv:2005.01643.

[87] LI, M., SHI, C., WU, Z. and FRYZLEWICZ, P. (2025). Testing stationarity and change point detection in reinforcement learning. *Ann. Statist.* **53**. MR4925122 https://doi.org/10.1214/25-aos2501

[88] LIAO, P., KLASNJA, P. and MURPHY, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *J. Amer. Statist. Assoc.* **116** 382–391. MR4227701 https://doi.org/10.1080/01621459.2020.1807993

[89] Liao, P., Klasnja, P. and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *J. Amer. Statist. Assoc.* **116** 382–391. MR4227701 https://doi.org/10.1080/01621459.2020.1807993

[90] Liao, P., Qi, Z., Wan, R., Klasnja, P. and Murphy, S. A. (2022). Batch policy learning in average reward Markov decision processes. *Ann. Statist.* **50** 3364–3387. MR4524500 https://doi.org/10.1214/22-aos2231

[91] Liu, Q., Li, L., Tang, Z. and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems* **31** 5356–5366.

[92] Liu, Y., Bacon, P.-L. and Brunskill, E. (2020). Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *Proceedings of the* 37*th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 6184–6193. PMLR.

[93] Lu, W., Zhang, H. H. and Zeng, D. (2013). Variable selection for optimal treatment decision. *Stat. Methods Med. Res.* **22** 493–504. MR3190671 https://doi.org/10.1177/0962280211428383

[94] Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919

[95] Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

[96] Luedtke, A. R. and van der Laan, M. J. (2017). Evaluating the impact of treating the optimal subgroup. *Stat. Methods Med. Res.* **26** 1630–1640. MR3687168 https://doi.org/10.1177/0962280217708664

[97] Mandel, T., Liu, Y., Levine, S., Brunskill, E. and Popovic, Z. (2014). Off-policy evaluation across representations with applications to educational games. In *Proceedings of the* 13*th International Conference on Autonomous Agentsand Multi-Agent Systems* 1077–1084.

[98] Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard Univ. Press, Cambridge, Mass.

[99] Marling, C. and Bunescu, R. C. (2018). The OhioT1DM dataset for blood glucose level prediction. In *KHD@ IJCAI*.

[100] Matsouaka, R. A., Li, J. and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics* **70** 489–499. MR3261768 https://doi.org/10.1111/biom.12179

[101] Meinshausen, N., Meier, L. and Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 https://doi.org/10.1198/jasa.2009.tm08647

[102] Miao, R., Qi, Z. and Zhang, X. (2022). Off-policy evaluation for episodic partially observable Markov decision processes under non-parametric models. arXiv preprint. Available at arXiv:2209.10064.

[103] Miao, W., Geng, Z. and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105** 987–993. MR3877879 https://doi.org/10.1093/biomet/asy038

[104] Miyaguchi, K. (2022). Almost hyperparameter-free hyperparameter selection framework for offline policy evaluation. In *AAAI Conference on Artificial Intelligence*.

[105] Mo, W., Qi, Z. and Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *J. Amer. Statist. Assoc.* **116** 659–674. MR4270012 https://doi.org/10.1080/01621459.2020.1796359

[106] Mou, W., Pananjady, A., Wainwright, M. J. and Bartlett, P. L. (2024). Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. *Math. Stat. Learn.* **7** 41–153. MR4733910 https://doi.org/10.4171/msl/44

[107] Mou, W., Wainwright, M. J. and Bartlett, P. L. (2022). Off-policy estimation of linear functionals: Non-asymptotic theory for semi-parametric efficiency. arXiv preprint. Available at arXiv:2209.13075.

[108] Mukherjee, R., Newey, W. K. and Robins, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. arXiv preprint. Available at arXiv:1705.07577.

[109] Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[110] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–355. MR1983752 https://doi.org/10.1111/1467-9868.00389

[111] Murphy, S. A., van der Laan, M. J. and Robins, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586 https://doi.org/10.1198/016214501753382327

[112] Nachum, O. and Dai, B. (2020). Reinforcement learning via Fenchel-Rockafellar duality. arXiv preprint. Available at arXiv:2001.01866.

[113] Namkoong, H., Keramati, R., Yadlowsky, S. and Brunskill, E. (2020). Off-policy policy evaluation for sequential decisions under unobserved confounding. arXiv preprint. Available at arXiv:2003.05623.

[114] Narita, Y., Yasui, S. and Yata, K. (2019). *Efficient Counterfactual Learning from Bandit Feedback*. AAAI Press, Menlo Park.

[115] Newey, W. K. (2013). Nonparametric instrumental variables estimation. *Amer. Econ. Rev.* **103** 550–56.

[116] Ning, Y., Peng, S. and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika* **107** 533–554. MR4138975 https://doi.org/10.1093/biomet/asaa020

[117] Owen, A. B. (2013). *Monte Carlo Theory, Methods and Examples*.

[118] Pananjady, A. and Wainwright, M. J. (2021). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Trans. Inf. Theory* **67** 566–585. MR4231973 https://doi.org/10.1109/TIT.2020.3027316

[119] Perdomo, J. C., Krishnamurthy, A., Bartlett, P. and Kakade, S. (2022). A Sharp Characterization of Linear Estimators for Offline Policy Evaluation. arXiv preprint. Available at arXiv:2203.04236.

[120] Precup, D., Sutton, R. and Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the* 17*th International Conference on Machine Learning* 759–766.

[121] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1270015

[122] Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

[123] Rashidinejad, P., Zhu, B., Ma, C., Jiao, J. and Russell, S. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory*

68 8156–8196. MR4544936 https://doi.org/10.1109/tit.2022.3185139
[124] ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable [MR2420458]. *Statist. Sci.* **22** 544–559. MR2420460 https://doi.org/10.1214/07-STS227D
[125] ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11
[126] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515
[127] ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* (*Minneapolis*, *MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681 https://doi.org/10.1007/978-1-4612-1284-3_1
[128] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730
[129] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118
[130] ROMANO, J. P. and DICICCIO, C. (2019). Multiple data splitting for testing. Dept. Statistics, Stanford Univ.
[131] ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2
[132] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41
[133] ROTNITZKY, A. and SMUCLER, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.* **21** Paper No. 188, 86. MR4209474
[134] RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880
[135] RUBIN, D. B. and VAN DER LAAN, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostat.* **4** Art. 5, 42. MR2399288 https://doi.org/10.2202/1557-4679.1084
[136] SALLAB, A. E., ABDOU, M., PEROT, E. and YOGAMANI, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electron. Imaging* **2017** 70–76.
[137] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. With comments and a rejoinder by the authors. MR1731478 https://doi.org/10.2307/2669923
[138] SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. MR3300363 https://doi.org/10.1214/13-STS450
[139] SHI, C., LU, W. and SONG, R. (2020). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *J. Mach. Learn. Res.* **21** Paper No. 176, 67. MR4209462
[140] SHI, C., SONG, R., LU, W. and FU, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 681–702. MR3849339 https://doi.org/10.1111/rssb.12273
[141] SHI, C., SONG, R., LU, W. and LI, R. (2021). Statistical inference for high-dimensional models via recursive online-score estimation. *J. Amer. Statist. Assoc.* **116** 1307–1318. MR4309274 https://doi.org/10.1080/01621459.2019.1710154
[142] SHI, C., UEHARA, M., HUANG, J. and JIANG, N. (2022). A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In *International Conference on Machine Learning* 20057–20094. PMLR.
[143] SHI, C., WAN, R., CHERNOZHUKOV, V. and SONG, R. (2021). Deeply-debiased off-policy interval estimation. arXiv preprint. Available at arXiv:2105.04646.
[144] SHI, C., WAN, R., SONG, R., LU, W. and LENG, L. (2020). Does the Markov decision process fit the data: Testing for the Markov property in sequential decision making. In *International Conference on Machine Learning* 8807–8817. PMLR.
[145] SHI, C., WANG, X., LUO, S., ZHU, H., YE, J. and SONG, R. (2023). Dynamic causal effects evaluation in A/B testing with a reinforcement learning framework. *J. Amer. Statist. Assoc.* **118** 2059–2071. MR4646626 https://doi.org/10.1080/01621459.2022.2027776
[146] SHI, C., ZHANG, S., LU, W. and SONG, R. (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 765–793. MR4460575
[147] SHI, C., ZHU, J., YE, S., LUO, S., ZHU, H. and SONG, R. (2024). Off-policy confidence interval estimation with confounded Markov decision process. *J. Amer. Statist. Assoc.* **119** 273–284. MR4713891 https://doi.org/10.1080/01621459.2022.2110878
[148] SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence—Volume* 2. *AAAI*'06 1219–1226. AAAI Press, Menlo Park.
[149] SI, N., ZHANG, F., ZHOU, Z. and BLANCHET, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proceedings of the* 37*th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 8884–8894. PMLR.
[150] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529** 484–489.
[151] SINGH, R. (2021). Debiased kernel methods. arXiv preprint. Available at arXiv:2102.11076.
[152] SMUCLER, E., SAPIENZA, F. and ROTNITZKY, A. (2022). Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika* **109** 49–65. MR4374640 https://doi.org/10.1093/biomet/asab018
[153] SONDHI, A., ARBOUR, D. and DIMMERY, D. (2020). Balanced off-policy evaluation in general action spaces. In *International Conference on Artificial Intelligence and Statistics* 2413–2423. PMLR.

[154] STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, Berlin.

[155] SU, Y., DIMAKOPOULOU, M., KRISHNAMURTHY, A. and DUDÍK, M. (2020). Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning* 9167–9176. PMLR.

[156] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[157] SWAMINATHAN, A. and JOACHIMS, T. (2015). The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems* **28** 3231–3239.

[158] SWAMINATHAN, A. and JOACHIMS, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Mach. Learn. Res.* **16** 1731–1755. MR3417797

[159] TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490 https://doi.org/10.1093/biomet/asq035

[160] TANG, Z., FENG, Y., LI, L., ZHOU, D. and LIU, Q. (2020). *Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation*. *ICLR* 2020.

[161] TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112

[162] TENNENHOLTZ, G., SHALIT, U. and MANNOR, S. (2020). Off-policy evaluation in partially observable environments. *Proc. AAAI Conf. Artif. Intell.* **34** 10276–10283.

[163] THOMAS, P. and BRUNSKILL, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the* 33*rd International Conference on Machine Learning* 2139–2148.

[164] TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence* 567–573. Amer. Assoc. of Artificial Intelligence, Menlo Park.

[165] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[166] TSIATIS, A. A. and DAVIDIAN, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data [MR2420458]. *Statist. Sci.* **22** 569–573. MR2420466 https://doi.org/10.1214/07-STS227B

[167] TSIATIS, A. A., DAVIDIAN, M. and CAO, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67** 536–545. MR2829022 https://doi.org/10.1111/j.1541-0420.2010.01476.x

[168] TSIATIS, A. A., DAVIDIAN, M., HOLLOWAY, S. T. and LABER, E. B. (2020). *Dynamic Treatment Regimes—Statistical Methods for Precision Medicine*. *Monographs on Statistics and Applied Probability* **164**. CRC Press, Boca Raton, FL. MR4920847

[169] UEHARA, M., HUANG, J. and JIANG, N. (2020). *Minimax Weight and Q-function Learning for Off-Policy Evaluation*. *ICML* 2020.

[170] UEHARA, M., IMAIZUMI, M., JIANG, N., KALLUS, N., SUN, W. and XIE, T. (2021). *Finite Sample Analysis of Minimax Offline Reinforcement Learning*: *Completeness*, *Fast Rates and First-Order Efficiency*.

[171] UEHARA, M., KIYOHARA, H., BENNETT, A., CHERNOZHUKOV, V., JIANG, N., KALLUS, N., SHI, C. and SUN, W. (2022). Future-Dependent Value-Based Off-Policy Evaluation in POMDPs. arXiv preprint. Available at arXiv:2207.13081.

[172] UEHARA, M. and SUN, W. (2022). *Pessimistic Model-Based Offline RL*: *PAC Bounds and Posterior Sampling Under Partial Coverage*. *ICLR* 2022.

[173] UENO, T., MAEDA, S., KAWANABE, M. and ISHII, S. (2011). Generalized TD learning. *J. Mach. Learn. Res.* **12** 1977–2020. MR2819024

[174] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. *Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

[175] VAN DER LAAN, M. J. and ROSE, S. (2018). *Targeted Learning in Data Science*: *Causal Inference for Complex Longitudinal Studies*. *Springer Series in Statistics*. Springer, Cham. MR3791826 https://doi.org/10.1007/978-3-319-65304-4

[176] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[177] VERMEULEN, K. and VANSTEELANDT, S. (2015). Bias-reduced doubly robust estimation. *J. Amer. Statist. Assoc.* **110** 1024–1036. MR3420681 https://doi.org/10.1080/01621459.2014.958155

[178] WAGER, S. and WALTHER, G. (2016). Adaptive concentration of regression trees, with application to random forests. arXiv preprint. Available at arXiv:1503.06388.

[179] WANG, L., ROTNITZKY, A., LIN, X., MILLIKAN, R. E. and THALL, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J. Amer. Statist. Assoc.* **107** 493–508. MR2980060 https://doi.org/10.1080/01621459.2011.641416

[180] WANG, L., ZHOU, Y., SONG, R. and SHERWOOD, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113** 1243–1254. MR3862354 https://doi.org/10.1080/01621459.2017.1330204

[181] WANG, R., FOSTER, D. P. and KAKADE, S. M. (2020). What are the statistical limits of offline RL with linear function approximation? arXiv preprint. Available at arXiv:2010.11895.

[182] WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

[183] WANG, Y.-X., AGARWAL, A. and DUDIK, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the* 34*th International Conference on Machine Learning* 3589–3597.

[184] WU, Y. and WANG, L. (2021). Resampling-based confidence intervals for model-free robust inference on optimal treatment regimes. *Biometrics* **77** 465–476. MR4307648 https://doi.org/10.1111/biom.13337

[185] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. arXiv preprint. Available at arXiv:2106.06926.

[186] XIE, T., MA, Y. and WANG, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems* **32** 9665–9675.

[187] XU, Z., LI, Z., GUAN, Q., ZHANG, D., LI, Q., NAN, J., LIU, C., BIAN, W. and YE, J. (2018). Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 905–913.

[188] YANG, M., NACHUM, O., DAI, B., LI, L. and SCHUURMANS, D. (2020). Off-policy evaluation via the regularized Lagrangian. arXiv preprint. Available at arXiv:2007.03438.

[189] YIN, M., BAI, Y. and WANG, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 1567–1575. PMLR.

[190] YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *Proceedings of the* 23*nd International Workshop on Artificial Intelligence and Statistics*.

[191] YOUNG, J. G., HERNÁN, M. A. and ROBINS, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiol*. *Methods* **3** 1–19.

[192] YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J. Y., LEVINE, S., FINN, C. and MA, T. (2020). MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems* **33** 14129–14142.

[193] ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

[194] ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694. MR3094445 https://doi.org/10.1093/biomet/ast014

[195] ZHANG, J. and BAREINBOIM, E. (2016). Markov decision processes with unobserved confounders: a causal approach. Technical Report R-23, Purdue AI Lab.

[196] ZHANG, J. and BAREINBOIM, E. (2021). Non-Parametric Methods for Partial Identification of Causal Effects. Columbia CausalAI Laboratory Technical Report (R-72).

[197] ZHANG, M. R., PAINE, T. L., NACHUM, O., PADURARU, C., TUCKER, G., WANG, Z. and NOROUZI, M. (2021). Autoregressive Dynamics Models for Offline Policy Evaluation and Optimization. arXiv preprint. Available at arXiv:2104.13877.

[198] ZHANG, R., ZHANG, X., NI, C. and WANG, M. (2022). Off-policy fitted Q-evaluation with differentiable function approximators: Z-estimation and inference theory. arXiv preprint. Available at arXiv:2202.04970.

[199] ZHANG, Y., LABER, E. B., DAVIDIAN, M. and TSIATIS, A. A. (2018). Interpretable dynamic treatment regimes. *J. Amer. Statist*. *Assoc*. **113** 1541–1549.

[200] ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895–904. MR3436715 https://doi.org/10.1111/biom.12354

[201] ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist*. *Assoc*. **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

[202] ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist*. *Assoc*. **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

[203] ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. *Springer Ser. Statist*. 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

[204] ZHOU, W., ZHU, R. and QU, A. (2024). Estimating optimal infinite horizon dynamic treatment regimes via pT-learning. *J. Amer. Statist*. *Assoc*. **119** 625–638. MR4713918 https://doi.org/10.1080/01621459.2022.2138760

[205] ZHU, R., ZHAO, Y.-Q., CHEN, G., MA, S. and ZHAO, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* **73** 391–400. MR3665956 https://doi.org/10.1111/biom.12593

[206] ZHU, W., ZENG, D. and SONG, R. (2019). Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes. *J. Amer. Statist*. *Assoc*. **114** 1404–1417. MR4011788 https://doi.org/10.1080/01621459.2018.1506341

Statistical Science
2026, Vol. 41, No. 3, 582–597
https://doi.org/10.1214/25-STS988

Gibbs Optimal Design of Experiments

Antony M. Overstall, Jacinta Holloway-Brown and James M. McGree

Abstract. Bayesian optimal design is a well-established approach to planning experiments. A distribution for the responses, that is, a statistical model, is assumed, which is dependent on unknown parameters. A utility function is then specified giving gain in information in estimating the true values of the parameters, using the Bayesian posterior distribution. A Bayesian optimal design is given by maximizing expectation of the utility with respect to the distribution implied by statistical model and prior distribution for the true parameter values. The approach accounts for the experimental aim, via specification of the utility, and of assumed sources of uncertainty. However, it is predicated on the statistical model being correct. Recently, a new type of statistical inference, known as Gibbs inference, has been proposed. This is Bayesian-like, that is, uncertainty for unknown quantities is represented by a posterior distribution, but does not necessarily require specification of a statistical model. The resulting inference is less sensitive to misspecification of the statistical model. This paper introduces Gibbs optimal design: a framework for optimal design of experiments under Gibbs inference. A computational approach to find designs in practice is outlined and the framework is demonstrated on exemplars including linear models, and experiments with count and time-to-event responses.

Key words and phrases: Gibbs statistical inference, loss function, robust design of experiments, utility function.

REFERENCES

ATKINSON, A. C., DONEV, A. N. and TOBIAS, R. D. (2007). *Optimum Experimental Designs, with SAS*. *Oxford Statistical Science Series* **34**. Oxford Univ. Press, Oxford. MR2323647

AZRIEL, D. (2023). Optimal minimax random designs for weighted least squares estimators. *Biometrika* **110** 273–280. MR4565457 https://doi.org/10.1093/biomet/asac016

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58. MR0189176 https://doi.org/10.1214/aoms/1177699477

BINGHAM, D. R. and CHIPMAN, H. A. (2007). Incorporating prior information in optimal design for model selection. *Technometrics* **49** 155–163. MR2380531 https://doi.org/10.1198/004017007000000038

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

BOX, G. E. P. and DRAPER, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654. MR0108872

CATONI, O. (2007). *Pac–Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

CHALONER, K. and LARNTZ, K. (1992). Bayesian design for accelerated life testing. *J. Statist. Plann. Inference* **33** 245–259.

CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304. MR1390519

COOK, R. D. and NACHTSHEIM, C. J. (1982). Model robust, linear-optimal designs. *Technometrics* **24** 49–54. MR0653111 https://doi.org/10.2307/1267577

DAVISON, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

ETZIONI, R. and KADANE, J. B. (1993). Optimal experimental design for another's analysis. *J. Amer. Statist. Assoc.* **88** 1404–1411. MR1245377

GILMOUR, S. G. and TRINCA, L. A. (2012). Optimum design of experiments for statistical inference. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **61** 345–401. MR2914518 https://doi.org/10.1111/j.1467-9876.2011.01000.x

Antony M. Overstall is an Associate Professor in the School of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom (e-mail: A.M.Overstall@soton.ac.uk). Jacinta Holloway-Brown is a Lecturer in the School of Computer and Mathematical Sciences, University of Adelaide, Adelaide 5005, Australia (e-mail: jacinta.holloway-brown@adelaide.edu.au). James M. McGree is a Professor in the School of Mathematical Sciences, Queensland University of Technology, Brisbane 4001, Australia (e-mail: james.mcgree@qut.edu.au).

GOOS, P. and JONES, B. (2011). *Optimal Design of Experiments*: *A Case Study Approach*. Wiley, New York.

GRÜNWALD, P. (2012). The safe Bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 169–183. Springer, Heidelberg. MR3042889 https://doi.org/10.1007/978-3-642-34106-9_16

HAYASHI, F. (2000). *Econometrics*. Princeton Univ. Press, Princeton, NJ. MR1881537

HOLMES, C. C. and WALKER, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104** 497–503. MR3698270 https://doi.org/10.1093/biomet/asx010

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). *Bayesian Survival Analysis*. *Springer Series in Statistics*. Springer, New York. MR1876598 https://doi.org/10.1007/978-1-4757-3447-8

JEWSON, J. and ROSSELL, D. (2022). General Bayesian loss function selection and the use of improper models. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **84** 1640–1665. MR4515553

JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380. MR3371010 https://doi.org/10.1093/biomet/asv002

KONSTANTINOU, M., BIEDERMANN, S. and KIMBER, A. (2017). Model robust designs for survival trials. *Comput*. *Statist*. *Data Anal*. **113** 239–250. MR3662404 https://doi.org/10.1016/j.csda.2016.10.013

KONSTANTINOU, M., BIEDERMANN, S. and KIMBER, A. C. (2015). Optimal designs for full and partial likelihood information—with application to survival models. *J*. *Statist*. *Plann*. *Inference* **165** 27–37. MR3350256 https://doi.org/10.1016/j.jspi.2015.03.007

KOTZ, S. and NADARAJAH, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge Univ. Press, Cambridge. MR2038227 https://doi.org/10.1017/CBO9780511550683

LI, W. and NACHTSHEIM, C. J. (2000). Model-robust factorial designs. *Technometrics* **42** 345–352.

LONG, Q., SCAVINO, M., TEMPONE, R. and WANG, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Comput*. *Methods Appl*. *Mech*. *Engrg*. **259** 24–39. MR3063259 https://doi.org/10.1016/j.cma.2013.02.017

LYDDON, S. P., HOLMES, C. C. and WALKER, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106** 465–478. MR3949315 https://doi.org/10.1093/biomet/asz006

MCGREE, J. M. and ECCLESTON, J. A. (2012). Robust designs for Poisson regression models. *Technometrics* **54** 64–72. MR2904736 https://doi.org/10.1080/00401706.2012.648867

MEAD, R., GILMOUR, S. G. and MEAD, A. (2012). *Statistical Principles for the Design of Experiments*: *Applications to Real Experiments*. Cambridge University Press, Cambridge.

MEYER, R. K. and NACHTSHEIM, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics* **37** 60–69. MR1322047 https://doi.org/10.2307/1269153

MORRIS, M. D. (2011). *Design of Experiments*: *An Introduction Based on Linear Models*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2682730

MYERS, R. H., MONTGOMERY, D. C., VINING, G. G. and ROBINSON, T. J. (2010). *Generalized Linear Models with Applications in Engineering and the Sciences*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2683129 https://doi.org/10.1002/9780470556986

O'HAGAN, A. and FORSTER, J. (2004). *Kendall's Advanced Theory of Statistics*. *Vol*. 2*B*. *Bayesian Inference*. Wiley, Chichester. MR3237119

OVERSTALL, A. and MCGREE, J. (2022). Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Anal*. **17** 1021–1041. MR4506020 https://doi.org/10.1214/21-ba1286

OVERSTALL, A. M., HOLLOWAY-BROWN, J. and MCGREE, J. M. (2026). Supplement to "Gibbs optimal design of experiments." https://doi.org/10.1214/25-STS988SUPP

OVERSTALL, A. M., MCGREE, J. M. and DROVANDI, C. C. (2018). An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Stat*. *Comput*. **28** 343–358. MR3747567 https://doi.org/10.1007/s11222-017-9734-x

OVERSTALL, A. M. and WOODS, D. C. (2017). Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59** 458–470. MR3740963 https://doi.org/10.1080/00401706.2016.1251495

OVERSTALL, A. M., WOODS, D. C. and ADAMOU, M. (2020). Acebayes: An R package for Bayesian optimal design of experiments via approximate coordinate exchange. *J*. *Stat*. *Softw*. **95** 1–33.

PAWITAN, Y. (2013). *In All Likelihood*. Oxford University Press, London.

RAINFORTH, T., FOSTER, A., IVANOVA, D. R. and BICKFORD SMITH, F. (2024). Modern Bayesian experimental design. *Statist*. *Sci*. **39** 100–114. MR4718529 https://doi.org/10.1214/23-sts915

ROBERT, C. P. (2007). *The Bayesian Choice*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2723361

RUSSELL, K. G. (2019). *Design of Experiments for Generalized Linear Models*. *Chapman & Hall/CRC Interdisciplinary Statistics Series*. CRC Press, Boca Raton, FL. MR3970305

RYAN, E. G., DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *Int*. *Stat*. *Rev*. **84** 128–154. MR3491282 https://doi.org/10.1111/insr.12107

SMUCKER, B. J. and DREW, N. M. (2015). Approximate model spaces for model-robust experiment design. *Technometrics* **57** 54–63. MR3318349 https://doi.org/10.1080/00401706.2014.897262

STIGLER, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard Univ. Press, Cambridge, MA. MR3585675 https://doi.org/10.4159/9780674970199

TSAI, P.-W. and GILMOUR, S. G. (2010). A general criterion for factorial designs under model uncertainty. *Technometrics* **52** 231–242. MR2757218 https://doi.org/10.1198/TECH.2010.08093

TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann*. *Statist*. **43** 2331–2352. MR3405596 https://doi.org/10.1214/15-AOS1314

WAITE, T. W. and WOODS, D. C. (2022). Minimax efficient random experimental design strategies with application to model-robust design for prediction. *J*. *Amer*. *Statist*. *Assoc*. **117** 1452–1465. MR4480724 https://doi.org/10.1080/01621459.2020.1863221

WALKER, S. G. (2013). Bayesian inference with misspecified models. *J*. *Statist*. *Plann*. *Inference* **143** 1621–1633. MR3082220 https://doi.org/10.1016/j.jspi.2013.05.013

WEISER, C. (2016). Methods for Multivariate Quadrature. mvQuad R package version 1.0-6.

WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. MR0909372 https://doi.org/10.1093/biomet/74.3.646

WIENS, D. P. (2015). Robustness of design. In *Handbook of Design and Analysis of Experiments* (A. Dean, M. D. Morris, J. Stufken and D. Bingham, eds.). *Chapman & Hall/CRC Handb*. *Mod*. *Stat*. *Methods* 719–753. CRC Press, Boca Raton, FL. MR3699366

WONG, R. K. W., STORLIE, C. B. and LEE, T. C. M. (2017). A frequentist approach to computer model calibration. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **79** 635–648. MR3611763 https://doi.org/10.1111/rssb.12182

Statistical Science
2026, Vol. 41, No. 3, 598–614
https://doi.org/10.1214/25-STS991

On Weighted Orthogonal Learners for Heterogeneous Treatment Effects

Paweł Morzywołek, Johan Decruyenaere and Stijn Vansteelandt

Abstract. Motivated by applications in personalized medicine and individualized policymaking, there is a growing interest in techniques for quantifying treatment effect heterogeneity in terms of the conditional average treatment effect (CATE). Some of the most prominent methods for CATE estimation developed in recent years are T-Learner, DR-Learner and R-Learner. The latter two were designed to improve on the former by being Neyman-orthogonal. However, the relations between them remain unclear, and likewise the literature remains vague on whether these learners converge to a useful quantity or (functional) estimand when the underlying optimization procedure is restricted to a class of functions that does not include the CATE. In this article, we provide insight into these questions by discussing DR-Learner and R-Learner as special cases of a general class of weighted Neyman-orthogonal learners for the CATE, for which we moreover derive oracle bounds. Our results shed light on how one may construct Neyman-orthogonal learners with desirable properties, on when DR-Learner may be preferred over R-Learner (and vice versa), and on novel learners that may sometimes be preferable to either of these. Theoretical findings are confirmed using results from simulation studies on synthetic data, as well as an application in critical care medicine.

Key words and phrases: Causal inference, heterogeneous treatment effects, orthogonal statistical learning, precision medicine.

REFERENCES

ATHEY, S. (2017). Beyond prediction: Using big data for policy problems. *Science* **355** 483–485.

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771 https://doi.org/10.3982/ECTA12723

BIBAUT, A., PETERSEN, M., VLASSIS, N., DIMAKOPOULOU, M. and VAN DER LAAN, M. (2021). Sequential causal inference in a single world of connected units. **1** 1–55.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHERNOZHUKOV, V., HANSEN, C., KALLUS, N., SPINDLER, M. and VASILIS, S. (2024). Applied Causal Inference Powered by ML and AI. 1–480.

COSTON, A., KENNEDY, E. and CHOULDECHOVA, A. (2020). Counterfactual predictions under runtime confounding. *Adv. Neural Inf. Process. Syst.* **33** 4150–4162.

CRUMP, R., HOTZ, J., IMBENS, G. and MITNIK, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Working Paper No. 330, National Bureau of Economic Research.

CURTH, A., ALAA, A. and VAN DER SCHAAR, M. (2020). Estimating structural target functions using machine learning and influence functions. arXiv preprint. Available at arXiv:2008.06461.

DAHABREH, I. J., ROBERTSON, S. E., STEINGRIMSSON, J. A., STUART, E. A. and HERNÁN, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Stat. Med.* **39** 1999–2014. MR4105273 https://doi.org/10.1002/sim.8426

FISHER, A. (2023). The connection between R-learning and inverse-

Paweł Morzywołek is a Postdoctoral Researcher, Department of Statistics, University of Washington, Seattle, Washington 98195-4322, USA (e-mail: pmorzywo@uw.edu). Johan Decruyenaere is the Professor of Intensive Care Medicine, Department of Intensive Care Medicine, Ghent University Hospital, Ghent, Belgium (e-mail: johan.decruyenaere@ugent.be). Stijn Vansteelandt is the Professor of Statistics, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium (e-mail: stijn.vansteelandt@ugent.be).

variance weighting for estimation of heterogeneous treatment effects. arXiv preprint. Available at arXiv:2307.09700.

FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258

HERNÁN, M., HSU, J. and HEALY, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance* **32** 42–49.

HERNÁN, M. and ROBINS, J. (2020). *Causal Inference—What If*. Chapman & Hall, Boca Raton, FL.

HINES, O., DUKES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2022). Demystifying statistical learning based on efficient influence functions. *Amer. Statist.* **76** 292–304. MR4453533 https://doi.org/10.1080/00031305.2021.2021984

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470. MR3086426 https://doi.org/10.1214/12-AOAS593

KALLUS, N. (2021). More efficient policy learning via optimal retargeting. *J. Amer. Statist. Assoc.* **116** 646–658. MR4270011 https://doi.org/10.1080/01621459.2020.1788948

KDIGO (2012). Kidney disease: Improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Inter., Suppl.* **2** 1–138.

KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

KÜNZEL, S. SEKHON, J., BICKEL, P. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165.

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12** 305–332. MR3505699 https://doi.org/10.1515/ijb-2015-0052

MCGRATH, S. and MUKHERJEE, R. (2022). On undersmoothing and sample splitting for estimating a doubly robust functional. arXiv preprint. Available at arXiv:2212.14857.

MENDELSON, S. and NEEMAN, J. (2010). Regularization in kernel learning. *Ann. Statist.* **38** 526–565. MR2590050 https://doi.org/10.1214/09-AOS728

MORZYWOŁEK, P., DECRUYENAERE, J. and VANSTEELANDT, S. (2026). Supplement to "On weighted orthogonal Learners for heterogeneous treatment effects." https://doi.org/10.1214/25-STS991SUPP

MORZYWOŁEK, P., STEEN, J., VANSTEELANDT, S., DECRUYENAERE, J., STERCKX, S. and VAN BIESEN, W. (2022). Timing of dialysis in acute kidney injury using routinely collected data and dynamic treatment regimes. *Crit. Care* **26** 1–11.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. *Rocz. Nauk Rol.* Tom **X** 1–51.

NEYMAN, J. (1979). $C(\alpha)$ tests and their use. *Sankhyā Ser. A* **41** 1–21. MR0615037

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

PFANZAGL, J. (1990). *Estimation in Semiparametric Models: Some Recent Developments. Lecture Notes in Statistics* **63**. Springer, New York. MR1048589 https://doi.org/10.1007/978-1-4612-3396-1

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **14** 923–945.

ROTNITZKY, A., SMUCLER, E. and ROBINS, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108** 231–238. MR4226201 https://doi.org/10.1093/biomet/asaa054

RUBIN, D. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

VAN GELOVEN, N., SWANSON, S., RAMSPEK, C., LUIJKEN, K., VAN DIEPEN, M., MORRIS, T., GROENWOLD, R., VAN HOUWELINGEN, H., PUTTER, H. et al. (2020). Prediction meets causal inference: The role of treatment in clinical prediction models. *Eur. J. Epidemiol.* **35** 619–630.

VANDERWEELE, T. J. and HERNAN, M. A. (2013). Causal inference under multiple versions of treatment. *J. Causal Inference* **1** 1–20. MR4289399 https://doi.org/10.1515/jci-2012-0002

VANSTEELANDT, S., MERTENS, K., SUETENS, C. and GOETGHEBEUR, E. (2009). Marginal structural models for partial exposure regimes. *Biostatistics* **10** 46–59.

VANSTEELANDT, S. and MORZYWOŁEK, P. (2023). Orthogonal prediction of counterfactual outcomes. arXiv preprint. Available at arXiv:2311.09423.

VAN DE GEER, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.* **8** 543–574. MR3211024 https://doi.org/10.1214/14-EJS894

VAN DER LAAN, L., CARONE, M. and LUEDTKE, A. (2024). Combining T-learning and DR-learning: a framework for oracle-efficient estimation of causal contrasts. arXiv preprint. Available at arXiv:2402.01972.

VAN DER LAAN, M. J. and LUEDTKE, A. R. (2014). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. Working Paper No. 332, U.C. Berkeley Division of Biostatistics Working Paper Series.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality. Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

Statistical Science
2026, Vol. 41, No. 3, 615–636
https://doi.org/10.1214/25-STS995

Restricted Maximum Likelihood Estimation in Generalized Linear Mixed Models

Luca Maestrini, Francis K.C. Hui and Alan H. Welsh

Abstract. Restricted maximum likelihood (REML) estimation is a widely accepted and frequently used method for fitting linear mixed models, with its principal advantage being that it produces less biased estimates of the variance components. However, the concept of REML does not immediately generalize to the setting of nonnormally distributed responses, and it is not always clear the extent to which, either asymptotically or in finite samples, such generalizations reduce the bias of variance component estimates compared to standard unrestricted maximum likelihood estimation. In this article, we review various attempts that have been made over the past four decades to extend REML estimation in generalized linear mixed models. We establish four major classes of approaches, namely approximate linearization, integrated likelihood, modified profile likelihoods and direct bias correction of the score function, and show that while these four classes may have differing motivations and derivations, they often arrive at a similar if not the same REML estimate. We compare the finite sample performance of these four classes, along with methods for REML estimation in hierarchical generalized linear models, through a numerical study involving binary and count data, with results demonstrating that all approaches perform similarly well reducing the finite sample size bias of variance components. Overall, we believe REML estimation should more widely adopted by practitioners using generalized linear mixed models, and that the exact choice of which REML approach to use should, at this point in time, be driven by software availability and ease of implementation.

Key words and phrases: Bias correction, modified profile likelihood, profile score function, random effects, variance estimation.

REFERENCES

ANDERSON, R. L. and BANCROFT, T. A. (1952). *Statistical Theory in Research*. McGraw-Hill Book Co, New York.

BAKBERGENULY, I. and KULINSKAYA, E. (2018). Meta-analysis of binary outcomes via generalized linear mixed models: A simulation study. *BMC Med. Res. Methodol.* **18** 1–18.

BARNDORFF-NIELSEN, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365. MR0712023 https://doi.org/10.1093/biomet/70.2.343

BELLHOUSE, D. R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika* **77** 743–746. MR1086686 https://doi.org/10.1093/biomet/77.4.743

BELLIO, R. and BRAZZALE, A. R. (2011). Restricted likelihood inference for generalized linear mixed models. *Stat. Comput.* **21** 173–183. MR2774850 https://doi.org/10.1007/s11222-009-9157-4

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200 https://doi.org/10.1214/ss/1009211803

BOLOGA, C. G., PANKRATZ, V. S., UNRUH, M. L., ROUMELIOTI, M. E., SHAH, V., SHAFFI, S. K., ARZHAN, S., COOK, J. and ARGYROPOULOS, C. (2021). High performance implementation of the hierarchical likelihood for generalized linear mixed models: An application to estimate the potassium reference range in massive electronic health records datasets. *BMC Med. Res. Methodol.* **21** 1–24.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

BRESLOW, N. E. and LIN, X. (1995). Bias correction in gener-

Luca Maestrini is a Lecturer, Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia (e-mail: luca.maestrini@anu.edu.au). Francis K.C. Hui is an Associate Professor, Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia (e-mail: francis.hui@anu.edu.au). Alan H. Welsh is an E.J. Hannan Professor of Statistics, Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia (e-mail: alan.welsh@anu.edu.au).

alised linear mixed models with a single component of dispersion. *Biometrika* **82** 81–91. MR1332840 https://doi.org/10.1093/biomet/82.1.81

BROOKS, M. E., KRISTENSEN, K., VAN BENTHEM, K. J., MAGNUSSON, A., BERG, C. W., NIELSEN, A., SKAUG, H. J., MAECHLER, M. and BOLKER, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9** 378–400.

COOPER, D. M. and THOMPSON, R. (1977). A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika* **64** 625–628. MR0488556 https://doi.org/10.2307/2345344

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B, Methodol.* **49** 1–39. With a discussion. MR0893334

COX, D. R. and REID, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79** 408–411. MR1185142 https://doi.org/10.1093/biomet/79.2.408

CRESSIE, N. and LAHIRI, S. N. (1993). The asymptotic distribution of REML estimators. *J. Multivariate Anal.* **45** 217–233. MR1221918 https://doi.org/10.1006/jmva.1993.1034

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103 https://doi.org/10.1214/aos/1018031103

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B, Methodol.* **39** 1–38. With discussion. MR0501537

DRUM, M. L. and MCCULLAGH, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49** 677–689. MR1243484 https://doi.org/10.2307/2532189

ELFF, M., HEISIG, J. P., SCHAEFFER, M. and SHIKANO, S. (2021). Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. *Br. J. Polit. Sci.* **51** 412–426.

ENGEL, B. and KEEN, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Stat. Neerl.* **48** 1–22. MR1267053 https://doi.org/10.1111/j.1467-9574.1994.tb01428.x

FELLNER, W. H. (1986). Robust estimation of variance components. *Technometrics* **28** 51–60. MR0824731 https://doi.org/10.2307/1269603

FELLNER, W. H. (1987). Sparse matrices, and the estimation of variance components by likelihood methods. *Comm. Statist. Simulation Comput.* **16** 439–463. MR0903158 https://doi.org/10.1080/03610918708812599

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GILMOUR, A. R., ANDERSON, R. D. and RAE, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72** 593–599. MR0817573 https://doi.org/10.1093/biomet/72.3.593

GOLDSTEIN, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **78** 45–51. MR1118229 https://doi.org/10.1093/biomet/78.1.45

GOLDSTEIN, H. and RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **159** 505–513. MR1413664 https://doi.org/10.2307/2983328

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Stat. Rev.* **55** 245–259. MR0963142 https://doi.org/10.2307/1403404

HALL, P., JOHNSTONE, I. M., ORMEROD, J. T., WAND, M. P. and YU, J. C. F. (2020). Fast and accurate binary response mixed model analysis via expectation propagation. *J. Amer. Statist. Assoc.* **115** 1902–1916. MR4189766 https://doi.org/10.1080/01621459.2019.1665529

HAN, J. and LEE, Y. (2024). Enhanced Laplace approximation. *J. Multivariate Anal.* **202** Paper No. 105321. MR4739369 https://doi.org/10.1016/j.jmva.2024.105321

HARVILLE, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61** 383–385. MR0368279 https://doi.org/10.1093/biomet/61.2.383

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340. MR0451550

HENDERSON, C. R. (1963). Selection index and expected genetic advance. In *Statisitical Genetics and Plant Breeding* 141–163. National Academy of Sciences, Washington DC.

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31** 423–447.

HUI, F. K. C. (2021). On the use of a penalized quasilikelihood information criterion for generalized linear mixed models. *Biometrika* **108** 353–365. MR4259136 https://doi.org/10.1093/biomet/asaa069

JIANG, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729. MR1631373 https://doi.org/10.2307/2670122

JIANG, J. (2013). The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *Ann. Statist.* **41** 177–195. MR3059414 https://doi.org/10.1214/13-AOS1084

JIANG, J., WAND, M. P. and BHASKARAN, A. (2022). Usable and precise asymptotics for generalized linear mixed model analysis and design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 55–82. MR4400390

JIN, S. and LEE, Y. (2021). A review of h-likelihood and hierarchical generalized linear model. *Wiley Interdiscip. Rev.: Comput. Stat.* **13** Paper No. e1527. MR4311088 https://doi.org/10.1002/wics.1527

JIN, S. and LEE, Y. (2024). Standard error estimates in hierarchical generalized linear models. *Comput. Statist. Data Anal.* **189** Paper No. 107852. MR4644814 https://doi.org/10.1016/j.csda.2023.107852

KENWARD, M. G. and ROGER, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53** 983–997.

KOSMIDIS, I. and FIRTH, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96** 793–804. MR2564491 https://doi.org/10.1093/biomet/asp055

KUK, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *J. Roy. Statist. Soc. Ser. B, Methodol.* **57** 395–407. MR1323346

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LEE, H. and LEE, Y. (2023). H-likelihood approach to deep neural networks with temporal-spatial random effects for high-cardinality categorical features. In *International Conference on Machine Learning* 18974–18987. PMLR.

LEE, W. and LEE, Y. (2012). Modifications of REML algorithm for HGLMs. *Stat. Comput.* **22** 959–966. MR2913795 https://doi.org/10.1007/s11222-011-9265-9

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B, Methodol.* **58** 619–678. With discussion. MR1410182

LEE, Y. and NELDER, J. A. (2001a). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88** 987–1006. MR1872215 https://doi.org/10.1093/biomet/88.4.987

LEE, Y. and NELDER, J. A. (2001b). Modelling and analysing correlated non-normal data. *Stat. Model.* **1** 3–16.

LEE, Y. and NELDER, J. A. (2003). Extended-REML estimators. *J. Appl. Stat.* **30** 845–856. MR2002126 https://doi.org/10.1080/0266476032000075930

LEE, Y., NELDER, J. A. and PAWITAN, Y. (2017). *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*, 2nd ed. Chapman and Hall/CRC.

LEE, Y. and NOH, M. (2018). dhglm: Double Hierarchical Generalized Linear Models. R package version 2.0.

LIANG, K.-Y., ZEGER, S. L. and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B, Methodol.* **54** 3–40. With discussion and a reply by the authors. MR1157713

LIAO, J. G. and LIPSITZ, S. R. (2002). A type of restricted maximum likelihood estimator of variance components in generalised linear mixed models. *Biometrika* **89** 401–409. MR1913967 https://doi.org/10.1093/biomet/89.2.401

LIN, L. and CHU, H. (2020). Meta-analysis of proportions using generalized linear mixed models. *Epidemiology* **31** 713–717.

LIN, X. and BRESLOW, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Amer. Statist. Assoc.* **91** 1007–1016. MR1424603 https://doi.org/10.2307/2291720

LINDSTROM, M. J. and BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46** 673–687. MR1085815 https://doi.org/10.2307/2532087

LOH, P.-R., BHATIA, G., GUSEV, A., FINUCANE, H. K., BULIK-SULLIVAN, B. K., POLLACK, S. J., SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM, DE CANDIA, T. R., LEE, S. H. et al. (2015a). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47** 1385–1392.

LOH, P.-R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJÁLMSSON, B. J., FINUCANE, H. K., SALEM, R. M., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M. et al. (2015b). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47** 284–290.

MAESTRINI, L., BHASKARAN, A. and WAND, M. P. (2024). Second term improvement to generalized linear mixed model asymptotics. *Biometrika* **111** 1077–1084. MR4794430 https://doi.org/10.1093/biomet/asad072

MAESTRINI, L., HUI, F. K. and WELSH, A. H. (2026). Supplement to "Restricted maximum likelihood estimation in generalized linear mixed models." https://doi.org/10.1214/25-STS995SUPPA, https://doi.org/10.1214/25-STS995SUPPB

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MCCULLAGH, P. and TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B, Methodol.* **52** 325–344. MR1064420

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89** 330–335.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105 https://doi.org/10.2307/2291460

MCCULLOCH, C. E. and SEARLE, S. R. (2004). *Generalized, Linear, and Mixed Models*. Wiley, New York.

MCGILCHRIST, C. A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc. Ser. B, Methodol.* **56** 61–69. MR1257795

MCGILCHRIST, C. A. and AISBETT, C. W. (1991). Restricted BLUP for mixed linear models. *Biom. J.* **33** 131–141. MR1128398 https://doi.org/10.1002/bimj.4710330202

MCGILCHRIST, C. A. and YAU, K. K. W. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Comm. Statist. Theory Methods* **24** 2963–2980. MR1364706 https://doi.org/10.1080/03610929508831663

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MCNEISH, D. (2019). Poisson multilevel models with small samples. *Multivar. Behav. Res.* **54** 444–455.

MEZA, C., JAFFRÉZIC, F. and FOULLEY, J.-L. (2009). Estimation in the probit normal model for binary outcomes using the SAEM algorithm. *Comput. Statist. Data Anal.* **53** 1350–1360. MR2657096 https://doi.org/10.1016/j.csda.2008.11.024

NING, X., HUI, F. K. C. and WELSH, A. H. (2026). Asymptotic results for penalized quasi-likelihood estimation in generalized linear mixed models. *Statist. Sinica* To appear.

NOH, M. and LEE, Y. (2007). REML estimation for binary data in GLMMs. *J. Multivariate Anal.* **98** 896–915. MR2325413 https://doi.org/10.1016/j.jmva.2006.11.009

NOH, M., LEE, Y., OUD, J. H. L. and TOHARUDIN, T. (2019). Hierarchical likelihood approach to non-Gaussian factor analysis. *J. Stat. Comput. Simul.* **89** 1555–1573. MR3934746 https://doi.org/10.1080/00949655.2019.1590575

ORMEROD, J. T. and WAND, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Statist.* **21** 2–17. MR2913353 https://doi.org/10.1198/jcgs.2011.09118

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554. MR0319325 https://doi.org/10.1093/biomet/58.3.545

PATTERSON, H. D. and THOMPSON, R. (1974). Maximum likelihood estimation of components of variance. In *Proceedings of the* 8*th International Biometric Conference* 197–207. Biometric Society, Washington DC.

PINHEIRO, J., BATES, D., DEBROY, S. and SARKAR, D. (2012). Nonlinear mixed-effects models. R Package Version 3.1-157 **3** 1–89.

PRENTICE, R. L. and ZHAO, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47** 825–839. MR1141951 https://doi.org/10.2307/2532642

R CORE TEAM (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

REICHERT, D. L. (1993). Variance component estimators for binary data derived from the dispersion-mean model. PhD thesis, Cornell University.

RICHARDSON, A. M. and WELSH, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Aust. J. Stat.* **36** 31–43. MR1309503 https://doi.org/10.1111/j.1467-842x.1994.tb00636.x

RODRÍGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **158** 73–89.

RUSSELL, T. S. and BRADLEY, R. A. (1958). One-way variances in a two-way classification. *Biometrika* **45** 111–129.

SAEI, A. and MCGILCHRIST, C. A. (1997). Random threshold models applied to inflated zero class data. *Aust. J. Stat.* **39** 5–16.

SAS Institute Inc. (2015). SAS/STAT 14.1 user's guide. SAS Institute Inc., Cary, NC.

SCHALL, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78** 719–727.

SEVERINI, T. A. (2000). *Likelihood Methods in Statistics*. *Oxford Statistical Science Series* **22**. Oxford Univ. Press, Oxford. MR1854870

SHUN, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. *J. Amer. Statist. Assoc.* **92** 341–349.

SHUN, Z. and MCCULLAGH, P. (1995). Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B, Methodol.* **57** 749–760. MR1354079

SMYTH, G. K. and VERBYLA, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. Roy. Statist. Soc. Ser. B, Methodol.* **58** 565–572. MR1394367

SMYTH, G. K. and VERBYLA, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* **10** 695–709.

STIRATELLI, R., LAIRD, N. and WARE, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40** 961–971.

STROUP, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Texts in Statistical Science Series.* CRC Press, Boca Raton, FL. MR2977489

TAWIAH, R., YAU, K. K. W., MCLACHLAN, G. J., CHAMBERS, S. K. and NG, S.-K. (2019). Multilevel model with random effects for clustered survival data with multiple failure outcomes. *Stat. Med.* **38** 1036–1055. MR3916714 https://doi.org/10.1002/sim.8041

THOMPSON, W. A. JR. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.* **33** 273–289. MR0133912 https://doi.org/10.1214/aoms/1177704731

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus*, 4th ed. *Statistics and Computing*. Springer, New York. MR1337030 https://doi.org/10.1007/978-1-4899-2819-1

VERBYLA, A. P. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Stat.* **32** 227–230.

VERBYLA, A. P. (2019). A note on model selection using information criteria for general linear models estimated using REML. *Aust. N. Z. J. Stat.* **61** 39–50. MR3942091 https://doi.org/10.1111/anzs.12254

WANG, K., YAU, K. K. W. and LEE, A. H. (2002). A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. *Stat. Med.* **21** 3639–3654.

WATSON, S. I. (2023). Generalised Linear Mixed Model Specification, Analysis, Fitting, and Optimal Design in R with the glmmr Packages. arXiv Preprint. Available at arXiv:2303.12657.

WOLFINGER, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80** 791–795. MR1282787 https://doi.org/10.1093/biomet/80.4.791

WOLFINGER, R. and O'CONNELL, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *J. Stat. Comput. Simul.* **48** 233–243.

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734 https://doi.org/10.1111/j.1467-9868.2010.00749.x

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R. Texts in Statistical Science Series.* CRC Press, Boca Raton, FL. MR3726911

YAU, K. and MCGILCHRIST, C. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Stat. Med.* **17** 1201–1213.

YAU, K. K. W. (2001). Multilevel models for survival analysis with random effects. *Biometrics* **57** 96–102. MR1833294 https://doi.org/10.1111/j.0006-341X.2001.00096.x

YAU, K. K. W. and KUK, A. Y. C. (2002). Robust estimation in generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 101–117. MR1881847 https://doi.org/10.1111/1467-9868.00327

YU, D. and YAU, K. K. W. (2013). Robust REML estimation for k-component Poisson mixture with random effects: Application to the epilepsy seizure count data and urinary tract infections data. *Stat. Med.* **32** 2479–2499. MR3067398 https://doi.org/10.1002/sim.5607

Statistical Science
2026, Vol. 41, No. 3, 637–658
https://doi.org/10.1214/25-STS996

High-Probability Minimax Lower Bounds

Tianyi Ma, Kabir A. Verchand and Richard J. Samworth

Abstract. The minimax risk is often considered as a gold standard against which we can compare specific statistical procedures. Nevertheless, as has been observed recently in robust and heavy-tailed estimation problems, the inherent reduction of the (random) loss to its expectation may entail a significant loss of information regarding its tail behavior. In an attempt to avoid such a loss, we introduce the notion of a minimax quantile, and seek to articulate its dependence on the quantile level. To this end, we develop high-probability variants of the classical Le Cam and Fano methods, as well as a technique to convert local minimax risk lower bounds to lower bounds on minimax quantiles. To illustrate the power of our framework, we deploy our techniques on several examples, recovering recent results in robust mean estimation and stochastic convex optimization, as well as obtaining several new results in covariance matrix estimation, sparse linear regression, nonparametric density estimation and isotonic regression. Our overall goal is to argue that minimax quantiles can provide a finer-grained understanding of the difficulty of statistical problems, and that, in wide generality, lower bounds on these quantities can be obtained via user-friendly tools.

Key words and phrases: Minimax quantiles, Le Cam's lemma, Fano's lemma.

REFERENCES

ADLER, R. J. and TAYLOR, J. E. (2009). *Random Fields and Geometry*. *Springer Monographs in Mathematics*. Springer, New York. MR2319516

ALIPRANTIS, C. D. and BORDER, K. C. (2006). *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, 3rd ed. Springer, Berlin. MR2378491

BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663 https://doi.org/10.1214/17-AOS1670

BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137. MR0523165 https://doi.org/10.1007/BF00535278

CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494 https://doi.org/10.1214/11-STS355

CANDÈS, E. J. and DAVENPORT, M. A. (2013). How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.* **34** 317–323. MR3008569 https://doi.org/10.1016/j.acha.2012.08.010

CARMON, Y. and HINDER, O. (2024). The price of adaptivity in stochastic convex optimization. arXiv preprint. Available at arXiv:2402.10898.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

CHEN, M., GAO, C. and REN, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Stat.* **10** 3752–3774. MR3579675 https://doi.org/10.1214/16-EJS1216

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

CHHOR, J., KLOPP, O. and TSYBAKOV, A. (2024). Generalized multiview model: Adaptive density estimation under low-rank constraints. arXiv preprint. Available at arXiv:2404.17209.

DAVIS, D., DRUSVYATSKIY, D., XIAO, L. and ZHANG, J. (2021). From low probability to high confidence in stochastic convex optimization. *J. Mach. Learn. Res.* **22** Paper No. 49, 38. MR4253742

Tianyi Ma is a PhD student at the Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK (e-mail: tm681@cam.ac.uk). Kabir A. Verchand is the Gary C. Butler Family Postdoctoral Fellow at the Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr NW, Atlanta, Georgia 30318, USA (e-mail: kverchand3@gatech.edu). Richard J. Samworth is the Professor of Statistical Science, Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK (e-mail: r.samworth@statslab.cam.ac.uk).

DEPERSIN, J. and LECUÉ, G. (2022a). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probab. Theory Related Fields* **183** 997–1025. MR4453320 https://doi.org/10.1007/s00440-022-01127-y

DEPERSIN, J. and LECUÉ, G. (2022b). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.* **50** 511–536. MR4382026 https://doi.org/10.1214/21-aos2118

DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. *Springer Series in Statistics*. Springer, New York. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082 https://doi.org/10.1214/aos/1176325367

DUCHI, J. C. (2018). Introductory lectures on stochastic optimization. In *The Mathematics of Data*. *IAS/Park City Math. Ser.* **25** 99–185. Amer. Math. Soc., Providence, RI. MR3839168

DUCHI, J. C. and HAQUE, S. (2024). An information-theoretic lower bound in time-uniform estimation. arXiv preprint. Available at arXiv:2402.08794.

EHRENFEUCHT, A., HAUSSLER, D., KEARNS, M. and VALIANT, L. (1989). A general lower bound on the number of examples needed for learning. *Inform. and Comput.* **82** 247–261. MR1016683 https://doi.org/10.1016/0890-5401(89)90002-3

GERCHINOVITZ, S., MÉNARD, P. and STOLTZ, G. (2020). Fano's inequality for random variables. *Statist. Sci.* **35** 178–201. MR4106600 https://doi.org/10.1214/19-STS716

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators*, *Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

HOPKINS, S. B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.* **48** 1193–1213. MR4102693 https://doi.org/10.1214/19-AOS1843

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

LECUÉ, G. and MENDELSON, S. (2013). Learning sub-Gaussian classes: Upper and minimax bounds. In *Topics in Learning Theory* Société Mathématique de France.

LUGOSI, G. and MENDELSON, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x

LUGOSI, G. and MENDELSON, S. (2019b). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

MA, T., VERCHAND, K. A., BERRETT, T. B., WANG, T. and SAMWORTH, R. J. (2024). Estimation beyond missing (completely) at random. arXiv preprint. Available at arXiv:2410.10704.

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson, a Wiley-Interscience Publication. MR0702836

ROBBINS, H. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* **62** 26–29. MR0069328 https://doi.org/10.2307/2308012

ROCKAFELLAR, R. T. (1997). *Convex Analysis*. *Princeton Landmarks in Mathematics*. Princeton Univ. Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks. MR1451876

SAMWORTH, R. J. and SEN, B. (2018). Editorial: Special issue on "Nonparametric inference under shape constraints". *Statist. Sci.* **33** 469–472. MR3881203 https://doi.org/10.1214/18-STS673

SAMWORTH, R. J. and SHAH, R. D. (2025). *Modern Statistical Methods and Theory*. Cambridge Univ. Press, Cambridge. To appear.

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459 https://doi.org/10.1007/s10208-011-9099-z

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B.C.*, 1974), *Vol.* 2 523–531. Canad. Math. Congr., Montreal, QC. MR0426989

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

WANG, H. and RAMDAS, A. (2025). Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance. *Sequential Anal.* **44** 56–110. MR4894312 https://doi.org/10.1080/07474946.2024.2428245

ZHIVOTOVSKIY, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electron. J. Probab.* **29** Paper No. 13, 28. MR4693860 https://doi.org/10.1214/23-ejp1021

ZHU, Z., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 2000–2031. MR4515564 https://doi.org/10.1111/rssb.12550

Statistical Science
2026, Vol. 41, No. 3, 659–670
https://doi.org/10.1214/25-STS997

The Fundamental Limits of Structure-Agnostic Functional Estimation

Sivaraman Balakrishnan, Edward Kennedy and Larry Wasserman

Abstract. Many recent developments in causal inference, and functional estimation problems more generally, have been motivated by the fact that classical one-step (first-order) debiasing methods, or their more recent sample-split double machine-learning avatars, can outperform plug-in estimators under surprisingly weak conditions. These first-order corrections improve on plug-in estimators in a black-box fashion, and consequently are often used in conjunction with powerful off-the-shelf estimation methods. On the other hand, these first-order methods are provably suboptimal in a minimax sense for functional estimation when the nuisance functions live in Hölder-type function spaces. This suboptimality of first-order debiasing has motivated the development of "higher-order" debiasing methods. The resulting estimators are, in some cases, provably optimal over Hölder-type spaces, but in sharp contrast to first-order estimators, both the estimators which are minimax-optimal and their analyses are crucially tied to properties of the underlying function space. Along a similar vein, some work has considered $\sqrt{n}$-consistent estimation of causal effects under weaker conditions than those required by first-order methods, once again relying on higher-order debiasing. More recent work in this area has focused on attempting to weaken the dependence of these higher-order estimators on the underlying nuisance function spaces, to make the resulting estimators and theory more robust. A central focus has been to try to make higher-order methods compatible with black-box nuisance estimators.

In this paper, we investigate the fundamental limits of structure-agnostic functional estimation, where relatively weak conditions are placed on the underlying nuisance functions. We show that there is a strong sense in which *existing first-order methods are optimal*. Particularly, we show that for several canonical integral functionals of interest it is impossible to improve on first-order estimators without making further, strong structural assumptions. We achieve this goal by providing a formalization of the problem of functional estimation with black-box nuisance function estimates, and deriving minimax lower bounds for this problem. Our results highlight some clear tradeoffs in functional estimation—if we wish to remain agnostic to the underlying nuisance function spaces, impose only high-level rate conditions, and maintain compatibility with black-box nuisance estimators then first-order methods are optimal. When we have a better understanding of the structure of the underlying nuisance functions then carefully constructed higher-order estimators can outperform first-order estimators.

Key words and phrases: Functional estimation, causal inference, double machine learning.

Sivaraman Balakrishnan is Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: the.seeing.stone@gmail.com). Edward Kennedy is Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, USA. Larry Wasserman is Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.

REFERENCES

[1] Balakrishnan, S., Kennedy, E. and Wasserman, L. (2026). Supplement to "The Fundamental Limits of Structure-Agnostic Functional Estimation." https://doi.org/10.1214/25-STS997SUPP

[2] Balakrishnan, S. and Wasserman, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

[3] Benkeser, D., Carone, M., van der Laan, M. J. and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

[4] Berrett, T. B. and Samworth, R. J. (2023). Efficient functional estimation and the super-oracle phenomenon. *Ann. Statist.* **51** 668–690. MR4600997 https://doi.org/10.1214/23-aos2265

[5] Berrett, T. B., Samworth, R. J. and Yuan, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.* **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

[6] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[7] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

[8] Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

[9] Cai, T. T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 391–419. MR4084169

[10] Cai, T. T., Levine, M. and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100** 126–136. MR2460482 https://doi.org/10.1016/j.jmva.2008.03.007

[11] Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346 https://doi.org/10.1214/10-AOS849

[12] Carpentier, A., Collier, O., Comminges, L., Tsybakov, A. B. and Wang, Y. (2022). Estimation of the ℓ_2-norm and testing in sparse linear regression with unknown variance. *Bernoulli* **28** 2744–2787. MR4474561 https://doi.org/10.3150/21-bej1436

[13] Cattaneo, M. D. and Jansson, M. (2022). Average density estimators: Efficiency and bootstrap consistency. *Econometric Theory* **38** 1140–1174. MR4524122 https://doi.org/10.1017/S0266466621000530

[14] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[15] Chernozhukov, V., Goldman, M., Semenova, V. and Taddy, M. (2017). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. arXiv preprint. Available at arXiv:1712.09988.

[16] Chernozhukov, V., Newey, W. K. and Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica* **90** 967–1027. MR4436042 https://doi.org/10.3982/ecta18515

[17] Collier, O., Comminges, L. and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444 https://doi.org/10.1214/15-AOS1432

[18] Comminges, L., Collier, O., Ndaoud, M. and Tsybakov, A. B. (2021). Adaptive robust estimation in sparse vector model. *Ann. Statist.* **49** 1347–1377. MR4298867 https://doi.org/10.1214/20-aos2002

[19] Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence. II, III. *Ann. Statist.* **19** 668–701. MR1105839 https://doi.org/10.1214/aos/1176348114

[20] Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043 https://doi.org/10.1016/0885-064X(90)90025-9

[21] Dukes, O., Vansteelandt, S. and Whitney, D. (2021). On doubly robust inference for double machine learning.

[22] Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294. MR1126325 https://doi.org/10.1214/aos/1176348249

[23] Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258

[24] Giné, E. and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli* **14** 47–61. MR2401653 https://doi.org/10.3150/07-BEJ110

[25] Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I: Theory of Statistics* 175–194. Univ. California Press, Berkeley, CA. MR0400513

[26] Hall, P. (1986). On powerful distributional tests based on sample spacings. *J. Multivariate Anal.* **19** 201–224. MR0853053 https://doi.org/10.1016/0047-259X(86)90027-8

[27] Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

[28] Ibragimov, I. A. and Khas'minskiĭ, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

[29] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446 https://doi.org/10.1007/978-0-387-21580-8

[30] Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **61** 2835–2885. MR3342309 https://doi.org/10.1109/TIT.2015.2412945

[31] Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L. and Robins, J. M. (2014). Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations.

[32] Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine* 231–261. CRC Press/CRC, Boca Raton.

[33] Kennedy, E. H., Balakrishnan, S., Robins, J. M. and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *Ann. Statist.* **52** 793–816. MR4744196 https://doi.org/10.1214/24-aos2369

[34] Kerkyacharian, G. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Ann.*

Statist. **24** 485–507. MR1394973 https://doi.org/10.1214/aos/1032894450
[35] KOZACHENKO, L. F. and LEONENKO, N. N. (1987). A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16. MR0908626
[36] KRISHNAMURTHY, A., KANDASAMY, K., PÓCZOS, B. and WASSERMAN, L. (2014). Nonparametric estimation of Rényi divergence and friends. In *Proceedings of the* 31*st International Conference on Machine Learning*.
[37] LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann*. *Statist*. **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458
[38] LE CAM, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, *Volume* 1: *Theory of Statistics*. Univ. California Press, Berkeley, CA.
[39] LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.
[40] LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020). On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statist*. *Sci*. **35** 518–539. MR4148230 https://doi.org/10.1214/20-STS786
[41] MANOLE, T., BALAKRISHNAN, S., NILES-WEED, J. and WASSERMAN, L. (2024). Plugin estimation of smooth optimal transport maps. *Ann*. *Statist*. **52** 966–998. MR4784066 https://doi.org/10.1214/24-aos2379
[42] MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. arXiv preprint. Available at arXiv:1705.07577.
[43] NEWEY, W. K. and ROBINS, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint. Available at arXiv:1801.09138.
[44] PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput*. **15** 1191–1253.
[45] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A*. *Freedman*. *Inst*. *Math*. *Stat*. (*IMS*) *Collect*. **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527
[46] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron*. *J*. *Stat*. **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479
[47] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann*. *Statist*. **34** 229–253. MR2275241 https://doi.org/10.1214/009053605000000877
[48] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann*. *Statist*. **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515
[49] ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48** 479–495. MR1173493 https://doi.org/10.2307/2532304
[50] ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J*. *Amer*. *Statist*. *Assoc*. **90** 122–129. MR1325119
[51] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J*. *Amer*. *Statist*. *Assoc*. **89** 846–866. MR1294730
[52] ROTNITZKY, A., SMUCLER, E. and ROBINS, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108** 231–238. MR4226201 https://doi.org/10.1093/biomet/asaa054
[53] SHEN, Y., GAO, C., WITTEN, D. and HAN, F. (2020). Optimal estimation of variance in nonparametric regression with random design. *Ann*. *Statist*. **48** 3589–3618. MR4185821 https://doi.org/10.1214/20-AOS1944
[54] SINGH, S. (2020). Estimating probability distributions and their properties PhD Thesis.
[55] SINGH, S., SRIPERUMBUDUR, B. K. and PÓCZOS, B. (2018). Minimax estimation of quadratic Fourier functionals. arXiv preprint. Available at arXiv:1803.11451.
[56] TCHETGEN, E., LI, L., ROBINS, J. and VAN DER VAART, A. (2008). Minimax estimation of the integral of a power of a density. *Statist*. *Probab*. *Lett*. **78** 3307–3311. MR2479495 https://doi.org/10.1016/j.spl.2008.07.001
[57] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926
[58] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794
[59] VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*'11—*Proceedings of the* 43*rd ACM Symposium on Theory of Computing* 685–694. ACM, New York. MR2932019 https://doi.org/10.1145/1993636.1993727
[60] VAN ES, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings. *Scand*. *J*. *Stat*. **19** 61–72. MR1172967
[61] VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int*. *J*. *Biostat*. **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038
[62] VAN DER VAART, A. (2014). Higher order tangent spaces and influence functions. *Statist*. *Sci*. **29** 679–686. MR3300365 https://doi.org/10.1214/14-STS478
[63] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
[64] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771
[65] WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann*. *Statist*. **36** 646–664. MR2396810 https://doi.org/10.1214/009053607000000901
[66] WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans*. *Inf*. *Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468

Statistical Science
2026, Vol. 41, No. 3, 671–696
https://doi.org/10.1214/25-STS1003

Sources of Uncertainty in Supervised Machine Learning—a Statisticians' View

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter and Göran Kauermann

Abstract. Supervised machine learning and predictive models have achieved an impressive standard today, enabling us to answer questions that were inconceivable a few years ago. Besides these successes, it becomes clear, that beyond pure prediction, which is the primary strength of most supervised machine learning algorithms, the quantification of uncertainty is relevant and necessary as well. However, before quantification is possible, types and sources of uncertainty need to be defined precisely. While first concepts and ideas in this direction have emerged in recent years, this paper adopts a conceptual, basic science perspective and examines possible sources of uncertainty. By adopting the viewpoint of a statistician, we discuss the concepts of aleatoric and epistemic uncertainty, which are more commonly associated with machine learning. The paper aims to formalize the two types of uncertainty and demonstrates that sources of uncertainty are miscellaneous and can not always be decomposed into aleatoric and epistemic. Drawing parallels between statistical concepts and uncertainty in machine learning, we emphasise the role of data and their influence on uncertainty.

Key words and phrases: Uncertainty, supervised machine learning, aleatoric uncertainty, epistemic uncertainty, label noise, measurement error, omitted variables, non-i.i.d. data, distribution shift, missing data, data-centric machine learning.

REFERENCES

ABDAR, M., POURPANAH, F., HUSSAIN, S., REZAZADEGAN, D., LIU, L., GHAVAMZADEH, M., FIEGUTH, P., CAO, X., KHOSRAVI, A. et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **76** 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

AGGARWAL, C. C., KONG, X., GU, Q., HAN, J. and PHILIP, S. Y. (2014). Active learning: A survey. In *Data Classification*: *Algorithms and Applications* 599–634. Chapman & Hall, London.

AKAIKE, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*. *Selected Papers of Hirotugu Akaike* 199–213.

ANDERSEN, R., KASPER, J. and FRANKEL, M. R. (1979). *Total Survey Error*: *Applications to Improve Health Surveys*, 1st ed. *The Jossey-Bass Social and Behavioral Science Series*. Jossey-Bass, San Francisco, CA.

AUGUSTIN, T., COOLEN, F. P. A., DE COOMAN, G. and TROFFAES, M. C. M. (2014). *Introduction to Imprecise Probabilities*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3236913 https://doi.org/10.1002/9781118763117

BARTLETT, P. L., MONTANARI, A. and RAKHLIN, A. (2021). Deep learning: A statistical viewpoint. *Acta Numer.* **30** 87–201. MR4295218 https://doi.org/10.1017/S0962492921000027

BARTLETT, P. L. and WEGKAMP, M. H. (2008). Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.* **9** 1823–1840. MR2438825

BATTLEDAY, R. M., PETERSON, J. C. and GRIFFITHS, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. Commun.* **11** 5418.

BECK, J., ECKMAN, S., MA, B., CHEW, R. and KREUTER, F. (2024). Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP* (*UncertaiNLP* 2024) (R. Vázquez, H. Celikkanat,

Cornelia Gruber is a Ph.D. Student, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany (e-mail: cornelia.gruber@stat.uni-muenchen.de). Patrick Oliver Schenk is a Ph.D. Student, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany (e-mail: patrick.schenk@stat.uni-muenchen.de). Malte Schierholz is Postdoctoral Researcher, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany (e-mail: malte.schierholz@stat.uni-muenchen.de). Frauke Kreuter is Professor, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany and Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: frauke.kreuter@stat.uni-muenchen.de). Göran Kauermann is Professor, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany (e-mail: goeran.kauermann@stat.uni-muenchen.de).

D. Ulmer, J. Tiedemann, S. Swayamdipta, W. Aziz, B. Plank, J. Baan and M.-C. de Marneffe, eds.) 81–86. Association for Computational Linguistics, St Julians, Malta.

Begoli, E., Bhattacharya, T. and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* **1** 20–23. https://doi.org/10.1038/s42256-018-0004-1

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

Bengs, V., Hüllermeier, E. and Waegeman, W. (2022). Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave and K. Cho, eds.).

Bothmann, L., Peters, K. and Bischl, B. (2025). What is fairness? On protected attributes and fictitious worlds. In *European Workshop on Algorithmic Fairness* 64–91. PMLR.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, IL.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A. and Syrgkanis, V. (2024). Long Story Short: Omitted Variable Bias in Causal Machine Learning. Available at arXiv:2112.13398 [econ]. https://doi.org/10.48550/arXiv.2112.13398

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15** 20170387. https://doi.org/10.1098/rsif.2017.0387

Clarté, L., Loureiro, B., Krzakala, F. and Zdeborová, L. (2022). A study of uncertainty quantification in overparametrized high-dimensional models.

Cressie, N. (1990). The origins of Kriging. *Math. Geol.* **22** 239–252. MR1047810 https://doi.org/10.1007/BF00889887

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. *Springer Texts in Statistics*. Springer, New York. MR1883764

Davison, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B. et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24** 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F. and Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the* 35*th International Conference on Machine Learning* 1184–1193. PMLR.

Dgani, Y., Greenspan, H. and Goldberger, J. (2018). Training a neural network based on unreliable human annotation of medical images. In 2018 *IEEE* 15*th International Symposium on Biomedical Imaging* (*ISBI* 2018) 39–42. IEEE.

Efron, B. (2020). Prediction, estimation, and attribution. *Int. Stat. Rev.* **88** S28–S59.

Egami, N. and Hartman, E. (2022). Elements of external validity: Framework, design, and analysis. *Amer. Polit. Sci. Rev.* **1**. https://doi.org/10.1017/S0003055422000880

El-Yaniv, R. and Wiener, Y. (2010). On the foundations of noise-free selective classification. *J. Mach. Learn. Res.* **11** 1605–1641. MR2653350

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. D. (2022). *Regression—Models, Methods and Applications*. Springer, Berlin. Corrected publication 2022. MR4404768 https://doi.org/10.1007/978-3-662-63882-8

Fan, J., Ma, C. and Zhong, Y. (2021). A selective overview of deep learning. *Statist. Sci.* **36** 264–290. MR4255194 https://doi.org/10.1214/20-sts783

Farquhar, S. and Gal, Y. (2022). What 'Out-of-distribution' Is and Is Not.

Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **25** 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Gal, Y. (2016). Uncertainty in deep learning. PhD thesis, Univ. Cambridge.

Gal, Y., Islam, R. and Ghahramani, Z. (2017). Deep Bayesian active learning with image data. In *Proceedings of the* 34*th International Conference on Machine Learning* 1183–1192. PMLR.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.* **46** 44:1–44:37. https://doi.org/10.1145/2523813

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P. et al. (2023). A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56** 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* **4** 1–58. Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA.

Geva, M., Goldberg, Y. and Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*) 1161–1166. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1107

Ghani, R. and Schierholz, M. (2020). Machine learning In *Big Data and Social Science*, 2nd ed. (I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter and J. Lane, eds.) **7**. CRC Press, Boca Raton, FL.

Girard, A., Rasmussen, C., Candela, J. Q. and Murray-Smith, R. (2002). Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems* **15**. MIT Press, Cambridge.

Goldstein, H. (2011). *Multilevel Statistical Models* 922. Wiley, New York.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

Groves, R. M. (2011a). 'Designed Data' and 'Organic Data'. US Census, Director's Blog, entry on 2011-05-31. Accessed 2022-07-27.

Groves, R. M. (2011b). Three eras of survey research. *Public Opin. Q.* **75** 861–871.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009). *Surv. Methodol.*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. MR3241261

GROVES, R. M. and LYBERG, L. (2010). Total survey error: Past, present, and future. *Public Opin*. *Q*. **74** 849–879. https://doi.org/10.1093/poq/nfq065

GRUBER, C., SCHENK, P. O., SCHIERHOLZ, M., KREUTER, F. and KAUERMANN, G. (2026). Supplement to "Sources of uncertainty in supervised machine learning—a statisticians' view." https://doi.org/10.1214/25-STS1003SUPP

HACKING, I. (1975). *The Emergence of Probability*: *A Philosophical Study of Early Ideas About Probability*, *Induction and Statistical Inference*. Cambridge Univ. Press, London. MR0467864

HAK, T. and BERNTS, T. (1996). Coder training: Theoretical training or practical socialization? *Qual*. *Sociol*. **19** 235–257. https://doi.org/10.1007/BF02393420

HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann*. *Statist*. **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HORA, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliab*. *Eng*. *Syst*. *Saf*. **54** 217–223. Treatment of Aleatory and Epistemic Uncertainty. https://doi.org/10.1016/S0951-8320(96)00077-4

HÜLLERMEIER, E. and WAEGEMAN, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach*. *Learn*. **110** 457–506. MR4234924 https://doi.org/10.1007/s10994-021-05946-3

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. and TAYLOR, J. (2023). *An Introduction to Statistical Learning—with Applications in Python*. *Springer Texts in Statistics*. Springer, Cham. MR4647520 https://doi.org/10.1007/978-3-031-38747-0

JAMES, G. M. (2003). Variance and bias for general loss functions. *Mach*. *Learn*. **51** 115–135. https://doi.org/10.1023/A:1022899518027

JAPEC, L., KREUTER, F., BERG, M., BIEMER, P. P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C. and USHER, A. (2015). Big data in survey research: AAPOR task force report. *Public Opin*. *Q*. **79** 839–880.

KAUERMANN, G., KÜCHENHOFF, H. and HEUMANN, C. (2021). *Statistical Foundations*, *Reasoning and Inference*. Springer, Berlin.

KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J*. *R*. *Stat*. *Soc*. *Ser*. *A* **179** 319–376. With discussion. MR3461587 https://doi.org/10.1111/rssa.12136

KEIDING, N. and LOUIS, T. A. (2018). Web-based enrollment and other types of self-selection in surveys and studies: Consequences for generalizability. *Annu*. *Rev*. *Stat*. *Appl*. **5** 25–47. MR3774738 https://doi.org/10.1146/annurev-statistics-031017-100127

KENDALL, A. and GAL, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Adv*. *Neural Inf*. *Process*. *Syst*. **30**.

KERN, C., ECKMAN, S., BECK, J., CHEW, R., MA, B. and KREUTER, F. (2023). Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics*: *EMNLP* 2023 (H. Bouamor, J. Pino and K. Bali, eds.) **992** 14874–14886. Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.findings-emnlp

KIUREGHIAN, A. D. and DITLEVSEN, O. (2009). Aleatory or epistemic? Does it matter? *Struct*. *Saf*. **31** 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020

KOMPA, B., SNOEK, J. and BEAM, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digit*. *Med*. **4**. https://doi.org/10.1038/s41746-020-00367-3

KREUTER, F., ed. (2013). *Improving Surveys with Paradata*: *Analytic Uses of Process Information* Wiley, Hoboken, NJ.

LAKSHMINARAYANAN, B., PRITZEL, A. and BLUNDELL, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) **30**. Curran Associates, Red Hook.

LEDERER, W. and KÜCHENHOFF, H. (2006). A short introduction to the SIMEX and MCSIMEX. *Newsl*. *R Proj*. **6** 26–31.

LEIBIG, C., ALLKEN, V., AYHAN, M. S., BERENS, P. and WAHL, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Sci*. *Rep*. **7** 1–14.

LI, Y. and IBANEZ-GUZMAN, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Process*. *Mag*. **37** 50–61. Conference Name: IEEE Signal Processing Magazine. https://doi.org/10.1109/MSP.2020.2973615

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LIU, H., ONG, Y.-S., SHEN, X. and CAI, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans*. *Neural Netw*. *Learn*. *Syst*. **31** 4405–4423. MR4169962 https://doi.org/10.1109/tnnls.2019.2957109

MALININ, A. and GALES, M. (2018). Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems* **31**. Curran Associates, Red Hook.

MASSING, N., WASMER, M., WOLF, C. and ZUELL, C. (2019). How standardized is occupational coding? A comparison of results from different coding agencies in Germany. *J*. *Off*. *Stat*. **35** 167–187. https://doi.org/10.2478/jos-2019-0008

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0907286

MCHUTCHON, A. and RASMUSSEN, C. (2011). Gaussian process training with input noise. In *Advances in Neural Information Processing Systems* **24**. Curran Associates, Red Hook.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K. and GALSTYAN, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput*. *Surv*. **54** 1–35.

MEI, S. and MONTANARI, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm*. *Pure Appl*. *Math*. **75** 667–766. MR4400901 https://doi.org/10.1002/cpa.22008

MISRA, I., LAWRENCE ZITNICK, C., MITCHELL, M. and GIRSHICK, R. (2016). Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 2930–2939.

MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2018). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. Second edition of [MR3057769]. MR3931734

MUKHOTI, J., KIRSCH, A., VAN AMERSFOORT, J., TORR, P. H. S. and GAL, Y. (2023). Deep deterministic uncertainty: A new simple baseline. In 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) 24384–24394. IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/CVPR52729.2023.02336

NAKKIRAN, P., KAPLUN, G., BANSAL, Y., YANG, T., BARAK, B. and SUTSKEVER, I. (2021). Deep double descent: Where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.* **12** Paper No. 124003, 32. Publisher: IOP Publishing and SISSA. MR4412831 https://doi.org/10.1088/1742-5468/ac3a74

NEAL, B., MITTAL, S., BARATIN, A., TANTIA, V., SCICLUNA, M., LACOSTE-JULIEN, S. and MITLIAGKAS, I. (2018). *A Modern Take on the Bias-Variance Tradeoff in Neural Networks*.

NEYSHABUR, B., LI, Z., BHOJANAPALLI, S., LECUN, Y. and SREBRO, N. (2018). Towards Understanding the Role of over-Parametrization in Generalization of Neural Networks.

NGUYEN, V.-L., SHAKER, M. H. and HÜLLERMEIER, E. (2022). How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn.* **111** 89–122. MR4375846 https://doi.org/10.1007/s10994-021-06003-9

NORTHCUTT, C. G., JIANG, L. and CHUANG, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *J. Artificial Intelligence Res.* **70** 1373–1411. MR4306614

PAN, Y., JING, Y., WU, T. and KONG, X. (2021). An integrated data and knowledge model addressing aleatory and epistemic uncertainty for oil condition monitoring. *Reliab. Eng. Syst. Saf.* **210** 107546. https://doi.org/10.1016/j.ress.2021.107546

PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595. MR3300360 https://doi.org/10.1214/14-STS486

PERDOMO, J. C., ZRNIC, T., MENDLER-DÜNNER, C. and HARDT, M. (2020). Performative Prediction. https://doi.org/10.48550/ARXIV.2002.06673

PETERSON, J. C., BATTLEDAY, R. M., GRIFFITHS, T. L. and RUSSAKOVSKY, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 9617–9626.

PSAROS, A. F., MENG, X., ZOU, Z., GUO, L. and KARNIADAKIS, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *J. Comput. Phys.* **477** Paper No. 111902, 83. MR4538412 https://doi.org/10.1016/j.jcp.2022.111902

QAYYUM, A., QADIR, J., BILAL, M. and AL-FUQAHA, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Rev. Biomed. Eng.* **14** 156–180. https://doi.org/10.1109/RBME.2020.3013489

QUIÑONERO-CANDELA, J., ed. (2009). *Dataset Shift in Machine Learning*. *Neural Information Processing Series* MIT Press, Cambridge, Mass. OCLC: ocn227205909.

RASMUSSEN, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass.

REN, J., LIU, P. J., FERTIG, E., SNOEK, J., POPLIN, R., DEPRISTO, M., DILLON, J. and LAKSHMINARAYANAN, B. (2019). Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems* **32**. Curran Associates, Red Hook.

REN, J., LUO, J., ZHAO, Y., KRISHNA, K., SALEH, M., LAKSHMINARAYANAN, B. and LIU, P. J. (2022). *Out-of-Distribution Detection and Selective Generation for Conditional Language Models*.

RODRÍGUEZ, P., BAUTISTA, M. A., GONZÀLEZ, J. and ESCALERA, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image Vis. Comput.* **75** 21–31.

ROLNICK, D., VEIT, A., BELONGIE, S. and SHAVIT, N. (2017). Deep learning is robust to massive label noise. arXiv preprint. Available at arXiv:1705.10694.

ROSSET, S. and TIBSHIRANI, R. J. (2020). From Fixed-X to Random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *J. Amer. Statist. Assoc.* **115** 138–151. Publisher: Taylor & Francis. MR4078450 https://doi.org/10.1080/01621459.2018.1424632

RUSSELL, S. J., NORVIG, P. and DAVIS, E. (2010). Artificial intelligence: A modern approach. In *Artificial Intelligence*, 3rd ed. *Prentice Hall Series in Artificial Intelligence*. Prentice Hall, Upper Saddle River, NJ.

SALLES, T., ROCHA, L. and GONÇALVES, M. (2021). A bias-variance analysis of state-of-the-art random forest text classifiers. *Adv. Data Anal. Classif.* **15** 379–405. MR4266383 https://doi.org/10.1007/s11634-020-00409-4

SCHENK, P. O., KERN, C. and BUSKIRK, T. D. (2025). Fares on fairness: Using a total error framework to examine the role of measurement and representation in training data on model fairness and bias. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness*, *EWAF'25*, *June 30th–July 2nd*, 2025, *Eindhoven*, *NL* (H. Weerts, M. Pechenizkiy, D. Allhutter, A. M. Corrêa, T. Grote and C. Liem, eds.) **294** 187–211. PMLR.

SCHENK, P. O. and REUSS, S. (2024). Paradata in surveys. In *Perspectives on Paradata*: *Research and Practice of Documenting Process Knowledge* (I. Huvila, L. Andersson and O. Sköld, eds.) 15–43. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-031-53946-6_2

SCHER, S. and MESSORI, G. (2018). Predicting weather forecast uncertainty with machine learning. *Q. J. R. Meteorol. Soc.* **144** 2830–2841. https://doi.org/10.1002/qj.3410

SCHERVISH, M. J. (1995). *Theory of Statistics*. *Springer Series in Statistics*. Springer, New York. MR1354146 https://doi.org/10.1007/978-1-4612-4250-5

SCHIERHOLZ, M. and SCHONLAU, M. (2020). Machine learning for occupation coding—a comparison study. *J. Surv. Stat. Methodol.* 1013–1034. https://doi.org/10.1093/jssam/smaa023

SCHWARTING, W., ALONSO-MORA, J. and RUS, D. (2018). Planning and decision-making for autonomous vehicles. *Annu. Rev. Control Robot. Auton. Syst.* **1** 187–210. https://doi.org/10.1146/annurev-control-060117-105157

SCHWEIGHOFER, K., AICHBERGER, L., IELANSKYI, M. and HOCHREITER, S. (2023). Introducing an Improved Information-Theoretic Measure of Predictive Uncertainty. arXiv Version Number: 1. https://doi.org/10.48550/ARXIV.2311.08309

SEEGER, M. (2004). Gaussian processes for machine learning. *Int. J. Neural Syst.* **14** 69–106.

SENGE, R., BÖSNER, S., DEMBCZYŃSKI, K., HAASENRITTER, J., HIRSCH, O., DONNER-BANZHOFF, N. and HÜLLERMEIER, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inform. Sci.* **255** 16–29. MR3117127 https://doi.org/10.1016/j.ins.2013.07.030

SHARMA, R., GARAYEV, H., KAUSHIK, M., PEIOUS, S. A., TIWARI, P. and DRAHEIM, D. (2022). Detecting Simpson's paradox: A machine learning perspective. In *Database and Expert Systems Applications*: 33*rd International Conference*, *DEXA* 2022, *Vienna*, *Austria*, *August* 22–24, 2022. *Proceedings*, *Part I* 323–335. Springer, Berlin.

SMITH, L. and GAL, Y. (2018). Understanding measures of uncertainty for adversarial example detection. arXiv preprint. Available at arXiv:1803.08533.

SULLIVAN, T. J. (2015). *Introduction to Uncertainty Quantification*. *Texts in Applied Mathematics* **63**. Springer, Cham. MR3364576 https://doi.org/10.1007/978-3-319-23395-6

VALDENEGRO-TORO, M. and SAROMO, D. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. In 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*) 1508–1516. IEEE, New Orleans, LA, USA. https://doi.org/10.1109/CVPRW56347.2022.00157

VALEN, J., BALKI, I., MENDEZ, M., QU, W., LEVMAN, J., BILBILY, A. and TYRRELL, P. N. (2022). Quantifying uncertainty in machine learning classifiers for medical imaging. *Internat. J. Comput. Assisted Radiol. Surg.* **17** 711–718. https://doi.org/10.1007/s11548-022-02578-3

VARSHNEY, K. R. (2022). *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA.

WAGNER, C. H. (1982). Simpson's paradox in real life. *Amer. Statist.* **36** 46–48.

WALLEY, P. (2000). Towards a unified theory of imprecise probability. *Internat. J. Approx. Reason.* **24** 125–148. MR1766280 https://doi.org/10.1016/S0888-613X(00)00031-1

WANG, W. and JING, B.-Y. (2022). Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *J. Mach. Learn. Res.* **23** Paper No. [193], 67. MR4577146

WEST, B. T., WAGNER, J., KIM, J. and BUSKIRK, T. D. (2023). The Total Data Quality Framework. Available at https://www.coursera.org/specializations/total-data-quality. Online, accessed 13 March 2023; cited figure is licensed under https://creativecommons.org/licenses/by-nc/4.0/ CC BY-NC 4.0.

WIMMER, L., SALE, Y., HOFMAN, P., BISCHL, B. and HÜLLERMEIER, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* 2282–2292. PMLR.

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. CRC Press, Boca Raton.

WOOLDRIDGE, J. M. (2005). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA. MR2768559

WRIGHT, W. A. (1999). Bayesian approach to neural-network modeling with input uncertainty. *IEEE Trans. Neural Netw.* **10** 1261–1270. Conference Name: IEEE Transactions on Neural Networks. https://doi.org/10.1109/72.809073

YANG, Z., YU, Y., YOU, C., STEINHARDT, J. and MA, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning* 10767–10777. PMLR.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64** 107–115. https://doi.org/10.1145/3446776

ZHOU, S. K., GREENSPAN, H., DAVATZIKOS, C., DUNCAN, J. S., VAN GINNEKEN, B., MADABHUSHI, A., PRINCE, J. L., RUECKERT, D. and SUMMERS, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE.*

ZHOU, X., LIU, H., POURPANAH, F., ZENG, T. and WANG, X. (2022). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing* **489** 449–465. https://doi.org/10.1016/j.neucom.2021.10.119

ZHU, X. X., HU, J., QIU, C., SHI, Y., KANG, J., MOU, L., BAGHERI, H., HABERLE, M., HUA, Y. et al. (2020). So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geosci. Remote Sens. Mag.* **8** 76–89.

ZHU, X. X., TUIA, D., MOU, L., XIA, G.-S., ZHANG, L., XU, F. and FRAUNDORFER, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **5** 8–36. https://doi.org/10.1109/MGRS.2017.2762307

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

IMS EDITORS